

# The error rate of learning halfspaces using Kernel-SVMs

Amit Daniely<sup>\*</sup>      Nati Linial<sup>†</sup>      Shai Shalev-Shwartz<sup>‡</sup>

November 9, 2012

## Abstract

We study the problem of (agnostically, efficiently and improperly) learning halfspaces with margin  $\gamma$ . Let  $\mathcal{D}$  be a distribution over labeled examples. The  $\gamma$ -margin error of a hyperplane  $h$  is the probability of an example to fall at the wrong side of  $h$  or at a distance of at most  $\gamma$  from it. The error of the best  $h$  is denoted  $\text{Err}_\gamma(\mathcal{D})$ . An efficient  $\alpha(\gamma)$ -approximation algorithm receives  $\gamma$  as input and in time  $\text{poly}(\frac{1}{\gamma})$ , using i.i.d. samples of  $\mathcal{D}$ , outputs a classifier with error rate  $\leq \alpha(\gamma) \text{Err}_\gamma(\mathcal{D})$ .

A popular family of such algorithms (E.g. Kernel SVM and Kernel ridge regression) operate by optimizing a convex loss over  $L^2$ -regularized linear class of functions. We show that, for every kernel and every convex loss, the approximation ratio of such an algorithm must be  $\geq \Omega\left(\frac{1/\gamma}{\text{poly}(\log(1/\gamma))}\right)$ . The result is tight, up to a poly-log factor.

## 1 Introduction

Support Vector Machine (SVM) is one of the most popular learning algorithms for the fundamental task of learning large margin halfspaces. SVM has been widely used in practice and its statistical efficiency has been thoroughly studied (Vapnik, 1998, Anthony and Bartlett, 1999, Schölkopf et al., 1998, Cristianini and Shawe-Taylor, 2000, Steinwart and Christmann, 2008). The basic variant of SVM, called Hard-SVM, assumes that the data is separable with a margin  $\gamma$ . That is, there exists a separating hyperplane that not only separates positive examples from negative examples but also has the property that all examples are at a distance of at least  $\gamma$  from it. Hard-SVM seeks a halfspace that separates the training examples with largest possible margin. Under the assumption of separability with margin, Hard-SVM is guaranteed to find a classifier whose classification error is at most  $\epsilon$  using time and sample complexity of  $\text{poly}(1/\gamma, 1/\epsilon)$ . However, this assumption is quite strong and rarely holds in practice. It is therefore natural to relax the problem and require instead that the algorithm learns a classifier whose error is comparable with that of the optimal  $\gamma$ -margin separating hyperplane.

---

<sup>\*</sup>Department of Mathematics, Hebrew University, Jerusalem 91904, Israel. amit.daniely@mail.huji.ac.il

<sup>†</sup>School of Computer Science and Engineering, Hebrew University, Jerusalem 91904, Israel. nati@cs.huji.ac.il

<sup>‡</sup>School of Computer Science and Engineering, Hebrew University, Jerusalem 91904, Israel. shais@cs.huji.ac.il

The second and most popular variant of SVM, called Soft-SVM, tackles the non-separable case. It solves a convex optimization problem by using the so-called hinge-loss function as a surrogate convex loss function to the number of margin violations. Replacing the number of margin violations with a surrogate convex loss is a heuristic. It is easy to see that this heuristic leads to an approximation ratio of  $\Theta(1/\gamma)$  (e.g. (Ben-David et al., 2012)). But, this lower bound is only known to hold when applying a surrogate convex loss function on the original representation of the examples. However, one of the most crucial reasons for the success of the SVM paradigm is the celebrated *kernel trick*. Kernels allow us to dramatically extend the repertoire of classification functions we can learn, by embedding the data in a high dimensional space. For example, using a polynomial kernel, we can learn all polynomials of bounded degree. This considerably extends the class of affine functions, used by the “straight” Soft-SVM. The goal of this paper is to derive strong lower bounds on the achievable approximation ratio of learning large margin halfspaces by applying SVM with *an arbitrary* kernel function. In addition, we derive a lower bound on the approximation ratio of any method that involves embedding of the examples into a *feature space* of dimension polynomial in  $1/\gamma$  and then learning a linear classifier in this feature space, by optimizing a convex surrogate.

## 1.1 Learning large margin halfspaces

Let  $B$  be the unit ball of some Hilbert space<sup>1</sup>  $H$ . A halfspace, parameterized by  $w \in B$  and  $b \in \mathbb{R}$ , is the classifier  $f(x) = \text{sign}(\Lambda_{w,b}(x))$ , where  $\Lambda_{w,b}(x) = \langle w, x \rangle + b$  is an affine functional. Given a distribution  $\mathcal{D}$  over  $B \times \{\pm 1\}$ , the error rate of  $\Lambda_{w,b}$  is

$$\text{Err}_{\mathcal{D},0-1}(\Lambda_{w,b}) = \Pr_{(x,y) \sim \mathcal{D}} (\text{sign}(\Lambda_{w,b}(x)) \neq y) = \Pr_{(x,y) \sim \mathcal{D}} (y\Lambda_{w,b}(x) \leq 0) .$$

More generally, we define the error rate of  $f : B \rightarrow \mathbb{R}$  as  $\text{Err}_{\mathcal{D},0-1}(f) = \Pr_{(x,y) \sim \mathcal{D}} (yf(x) \leq 0)$ . The  $\gamma$ -margin error rate of  $f$  is

$$\text{Err}_{\mathcal{D},\gamma}(f) = \Pr_{(x,y) \sim \mathcal{D}} (yf(x) \leq \gamma) .$$

Note that for a halfspace classifier  $\Lambda_{w,b}$ , if  $\|w\| = 1$  then  $|\Lambda_{w,b}(x)|$  is the distance of  $x$  from the separating hyperplane. Therefore, the  $\gamma$ -margin error rate is the probability of  $x$  to either be in the wrong side of the hyperplane or to be at a distance of at most  $\gamma$  from the hyperplane. The least  $\gamma$ -margin error rate of a halfspace classifier is denoted  $\text{Err}_{\gamma}(\mathcal{D}) = \min_{w \in B, b \in \mathbb{R}} \text{Err}_{\mathcal{D},\gamma}(\Lambda_{w,b})$ .

A learning algorithm receives  $\gamma, \epsilon$  and can receive i.i.d. samples from  $\mathcal{D}$ . The algorithm should return a classifier (which need not be an affine function). We say that the algorithm has approximation ratio  $\alpha(\gamma)$  if for every distribution, it outputs (w.h.p. over the i.i.d.  $\mathcal{D}$ -samples) a classifier with error rate  $\leq \alpha(\gamma) \text{Err}_{\gamma}(\mathcal{D}) + \epsilon$ . An *efficient* algorithm uses  $\text{poly}(1/\gamma, 1/\epsilon)$  time and samples, and the output classifier can be evaluated in  $\text{poly}(1/\gamma, 1/\epsilon)$  time.

---

<sup>1</sup>For concreteness, we assume that the Hilbert space is  $\ell^2$ . We will later on consider finite dimensional spaces by viewing  $\mathbb{R}^d$  as embedded in  $\ell^2$  by the standard embedding.

It is well known that for data which is separable with margin  $\gamma$ , i.e.  $\text{Err}_\gamma(\mathcal{D}) = 0$ , the Support Vector Machine (SVM) algorithm finds a classifier with error  $\leq \epsilon$  with time and sample complexity  $\leq \text{poly}(1/\gamma, 1/\epsilon)$ . The problem of proper<sup>2</sup> learning of halfspaces in the non-separable case without the margin assumption was shown to be hard to approximate for any constant approximation factor (Feldman et al., 2006, Guruswami and Raghavendra, 2006). Proper learning with the margin assumption was also shown to be hard (Ben-David and Simon, 2000). It has been recently shown that improper learning under the margin assumption is also hard (under some cryptographic assumptions), namely, no polynomial time algorithm can achieve an approximation ratio of  $\alpha(\gamma) = 1$ . As we note next, this is a far cry from what currently known algorithms can accomplish.

The best currently known efficient algorithms (Birnbaum and Shalev-Shwartz, 2012) and (Long and Servedio, 2011) achieve an approximation ratio  $\frac{1}{\gamma \cdot \sqrt{\log(1/\gamma)}}$ . The algorithm of (Birnbaum and Shalev-Shwartz, 2012) applies Soft-SVM with a particular kernel. Several other authors use similar techniques: Mapping the examples into some feature space and learning a halfspace there using a surrogate convex loss. For example, (Kalai et al., 2005, Blais et al., 2008) developed a PTAS for agnostically learning halfspaces under several distributional assumptions based on a polynomial transformation. This relies on embeddings that turn affine classifiers to polynomials of a certain degree. Our lower bounds apply to these polynomial spaces and to every finite dimensional space of classification functions.

The learning theory literature contains consistency results for learning with the so-called universal kernels and well-calibrated surrogate loss functions. This includes the study of asymptotic relations between surrogate convex loss functions and the 0-1 loss function (Zhang, 2004, Bartlett et al., 2006, Steinwart and Christmann, 2008). It is shown that the approximation ratio of SVM with a universal kernel goes to 1 as the sample size grows. Our result implies that this convergence is very slow, e.g., an exponentially large (in  $\frac{1}{\gamma}$ ) sample is needed to make the error  $< 2 \text{Err}_\gamma(\mathcal{D})$ .

Before stating our results, let us recall the kernel SVM approach as well as other feature mapping based approaches that do not depend on the kernel trick.

## 1.2 Kernel based learning and kernel SVM

The significance of separation with margin is a fundamental discovery of statistical learning theory. The affine function that minimizes the *empirical*  $\gamma$ -margin error rate over an i.i.d. sample of size  $\text{poly}(1/\gamma, 1/\epsilon)$  has error rate  $\leq \text{Err}_\gamma(\mathcal{D}) + \epsilon$ . However, this minimization problem is computationally hard (see (Ben-David and Simon, 2000)).

The SVM paradigm deals with this hardness by replacing the margin error rate with a *convex surrogate loss*, in particular, the *hinge loss*  $l_{\text{hinge}}(x) = (1 - x)_+$ . For simplicity, let us also define the 0 – 1 loss function,  $l_{0-1}(x) = \mathbf{1}(x \leq 0)$ , and the  $\gamma$ -margin loss function,  $l_\gamma(x) = \mathbf{1}(x \leq \gamma)$ . For any loss function,  $l$ , define  $\text{Err}_{\mathcal{D},l}(f) = \mathbb{E}_{(x,y) \sim \mathcal{D}}(l(yf(x)))$ . We also use shorthands such as  $\text{Err}_{\mathcal{D},\text{hinge}}$  instead of  $\text{Err}_{\mathcal{D},l_{\text{hinge}}}$ . Note that for  $x \in [-2, 2]$ ,

$$l_{0-1}(x) \leq l_{\text{hinge}}(x/\gamma) \leq (1 + 2/\gamma)l_\gamma(x) ,$$

---

<sup>2</sup>A proper learner must output a halfspace classifier. Here we consider improper learning where the learner can output any classifier.

from which it easily follows that a solution to the following problem

$$\min_{w,b} \text{Err}_{\mathcal{D},\text{hinge}} \left( \frac{1}{\gamma} \Lambda_{w,b} \right) \quad \text{s.t.} \quad w \in H, \quad b \in \mathbb{R}, \quad \|w\|_H \leq 1$$

gives an approximation ratio of  $\alpha(\gamma) = 1 + 2/\gamma$ . It is more convenient to use the problem

$$\min_{w,b} \text{Err}_{\mathcal{D},\text{hinge}} (\Lambda_{w,b}) \quad \text{s.t.} \quad w \in H, \quad b \in \mathbb{R}, \quad \|w\|_H \leq C, \quad (1)$$

which is equivalent for  $C = \frac{1}{\gamma}$ . Problem (1) can be approximated, up to an additive error of  $\epsilon$ , using a  $\text{poly}(\frac{1}{\gamma}, \frac{1}{\epsilon})$  sample and time.

Kernel-free soft SVM means solving Problem (1). Kernel SVM has the additional freedom of embedding  $B$  into the unit ball of another Hilbert space, thus extending the repertoire of possible output functions. Concretely, let  $\psi : B \rightarrow B_1$ , where  $B_1$  is the unit ball of a Hilbert space  $H_1$ . The embedding  $\psi$  need not be computed directly. Rather, it is enough that we can compute the corresponding *kernel*,  $k(x, y) := \langle \psi(x), \psi(y) \rangle_{H_1}$ . It remains to solve the following program

$$\min_{w,b} \text{Err}_{\mathcal{D},\text{hinge}} (\Lambda_{w,b} \circ \psi) \quad \text{s.t.} \quad w \in H_1, \quad b \in \mathbb{R}, \quad \|w\|_{H_1} \leq C. \quad (2)$$

This problem can be approximated, up to an additive error of  $\epsilon$ , using  $\text{poly}(C/\epsilon)$  samples and time as follows. Let  $(x_1, y_1), \dots, (x_n, y_n)$  be a sequence of i.i.d. samples from  $\mathcal{D}$ . Let  $\hat{D}$  be the empirical distribution over these examples. By a uniform convergence argument (e.g. (Boucheron et al., 2005)), if  $n = \Omega(C^2/\epsilon^2)$ , then w.h.p. over the choice of examples we have

$$\max_{b \in \mathbb{R}, w \in H_1, \|w\|_{H_1} \leq C} |\text{Err}_{\mathcal{D},\text{hinge}} (\Lambda_{w,b} \circ \psi) - \text{Err}_{\hat{D},\text{hinge}} (\Lambda_{w,b} \circ \psi)| \leq \epsilon/2.$$

Therefore, we can solve (2) w.r.t.  $\hat{D}$  instead of w.r.t.  $\mathcal{D}$ . It is easy to verify that there exists a solution of the form  $w = \sum_{i=1}^n \alpha_i \psi(x_i)$  to the problem w.r.t.  $\hat{D}$  for some real  $\alpha_i$ 's. Therefore, we can optimize over  $\alpha \in \mathbb{R}^n$  instead of over  $w \in H_1$ , which yields the problem

$$\min_{\alpha \in \mathbb{R}^n, b \in \mathbb{R}} \frac{1}{n} \sum_{j=1}^n l_{\text{hinge}} \left( y_j \left( \sum_{i=1}^n \alpha_i k(x_i, x_j) + b \right) \right) \quad \text{s.t.} \quad \sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j) \leq C.$$

In this formulation we access examples only via the kernel function (a.k.a. the “kernel trick”). This is a convex optimization problem on  $n+1$  variables, which can be solved in time  $\text{poly}(n)$  by standard methods (assuming that the kernel function can be evaluated efficiently). Recall that  $n$  should be  $\Omega(C^2/\epsilon^2)$  for the uniform convergence to hold. It therefore makes sense, and we will adopt this practice, to refer to  $C$  as the time and sample complexity of program (2). In particular, we will prove lower bounds to any approximate solution of program (2) assuming that  $C = \text{poly}(1/\gamma)$ .

We formulate our results for Problem (2), but they apply as well to the following commonly used formulation of the kernel SVM problem, where the constraint  $\|w\|_{H_1} \leq C$  is replaced by a regularization term. Namely

$$\min_{w \in H_1, b \in \mathbb{R}} \frac{1}{C^2} \|w\|_{H_1}^2 + \text{Err}_{\mathcal{D},\text{hinge}} (\Lambda_{w,b} \circ \psi) \quad (3)$$

The optimum of program (3) is  $\leq 1$  as the zero solution  $w = 0, b = 0$  shows. Thus, if  $w, b$  is an approximate optimal solution, then  $\frac{\|w\|_{H_1}^2}{C^2} \leq O(1) \Rightarrow \|w\|_{H_1} \leq O(C)$ . This observation makes it easy to modify our results on program (2) to apply to program (3).

### 1.3 Non Kernel-based Feature Mappings

The kernel trick constrains the  $L^2$ -norm of  $w \in H_1$ . Other algorithms require instead that  $w$  belongs to some set  $W \subseteq \mathbb{R}^m$ . For example, the Lasso method (Tibshirani, 1996) constrains  $w$  to have a low  $L^1$ -norm. We take the dimension of  $H_1$  as a measure of the complexity of the problem, and require it to be polynomial in  $1/\gamma$ . This choice is justified, as without the  $L^2$ -norm constraint, we cannot rely on the “kernel-trick” and in general must work directly in  $H_1$ . Therefore, every algorithm must have time complexity  $\Omega(m)$ . We prove lower bounds for any approximate solution to a problem of the form

$$\min_{w,b} \text{Err}_{\mathcal{D},l}(\Lambda_{w,b} \circ \psi) \quad \text{s.t.} \quad w \in W \subset \mathbb{R}^m, b \in \mathbb{R}, \quad (4)$$

where  $l$  is some surrogate loss function (see formal definition in the next section).

Note that any finite dimensional space of functions over the ball, that includes the constant functions, is of the form  $\{\Lambda_{w,b} \circ \psi\}$  for some  $\psi$ . Hence, our lower bounds hold for any method that optimizes a surrogate loss over a subset of a finite dimensional space of functions that includes the constant functions.

## 2 Results

We first define the two families of algorithms to which our lower bounds apply. We start with the class of *surrogate loss* functions. This class includes the most popular choices such as the absolute loss  $|1 - x|$ , the squared loss  $(1 - x)^2$ , the logistic loss  $\log_2(1 + e^{-x})$  etc.

**Definition 2.1 (Surrogate loss function)** *A function  $l : \mathbb{R} \rightarrow \mathbb{R}$  is called a surrogate loss function if  $l$  is convex and is lower bounded by the 0-1 loss.*

The first family of algorithms contains kernel based algorithms, such as kernel SVM.

**Definition 2.2 (Kernel based learner)** *Let  $l : \mathbb{R} \rightarrow \mathbb{R}$  be a surrogate loss function. A kernel based learning algorithm,  $A$ , receives as input  $\gamma \in (0, 1)$ . It then selects  $C = C_A(\gamma)$  and an absolutely continuous feature mapping,  $\psi = \psi_A(\gamma)$ . The algorithm returns a function*

$$A(\gamma) \in \{\Lambda_{w,b} \circ \psi : w \in H_1, b \in \mathbb{R}, \|w\|_{H_1} \leq C\}$$

*such that, with probability  $\geq 1 - \exp(-1/\gamma)$ ,*

$$\text{Err}_{\mathcal{D},l}(A(\gamma)) \leq \inf\{\text{Err}_{\mathcal{D},l}(\Lambda_{w,b} \circ \psi) : w \in H_1, b \in \mathbb{R}, \|w\|_{H_1} \leq C\} + \sqrt{\gamma}.$$

*We say that  $A$  is efficient if  $C_A(\gamma) \leq \text{poly}(1/\gamma)$ .*

Note that the definition of kernel based learner allows for any predefined convex surrogate loss, not just the hinge loss. Namely, we consider the program

$$\min_{w,b} \text{Err}_{\mathcal{D},l}(\Lambda_{w,b} \circ \psi) \quad \text{s.t.} \quad w \in H_1, b \in \mathbb{R}, \|w\|_{H_1} \leq C. \quad (5)$$

Uniform convergence results (e.g. (Boucheron et al., 2005)) suggest that for Lipschitz surrogate, program (5) can be approximated, up to an additive error of  $\epsilon$  and with probability  $> 1 - \exp(-1/\rho)$ , from a sample of size<sup>3</sup>  $\text{poly}(C, \frac{1}{\epsilon}, \frac{1}{\rho})$ , hence our definition of efficiency here. We note that our results also hold if the kernel corresponds to  $\psi$  is hard to compute.

The second family of learning algorithms involves an arbitrary feature mapping and domain constraint on the vector  $w$ , as in program (4).

**Definition 2.3 (Generalized linear learner)** *Let  $l : \mathbb{R} \rightarrow \mathbb{R}$  be some surrogate loss function. A generalized linear learning algorithm,  $A$ , receives as input  $\gamma \in (0, 1)$ . It then selects a continuous embedding  $\psi = \psi_A(\gamma) : B \rightarrow \mathbb{R}^m$  and a constraint set  $W = W_A(\gamma) \subseteq \mathbb{R}^m$ . The algorithm returns, with probability  $\geq 1 - \exp(1/\gamma)$ , a function*

$$A(\gamma) \in \{\Lambda_{w,b} \circ \psi : w \in W, b \in \mathbb{R}\}$$

such that

$$\text{Err}_{\mathcal{D},l}(A(\gamma)) \leq \inf\{\text{Err}_{\mathcal{D},l}(\Lambda_{w,b} \circ \psi) : w \in W, b \in \mathbb{R}\} + \sqrt{\gamma}.$$

We say that  $A$  is efficient if  $m = m_A(\gamma) \leq \text{poly}(1/\gamma)$ .

## 2.1 Main Results

**Theorem 2.4** *Let  $A$  be an efficient kernel-based learner. Then, for every  $\gamma > 0$ , there exists a distribution  $\mathcal{D}$  on  $B$  such that, w.p.  $\geq 1 - \exp(1/\gamma)$ ,*

$$\frac{\text{Err}_{\mathcal{D},0-1}(A(\gamma))}{\text{Err}_{\gamma}(\mathcal{D})} \geq \Omega\left(\frac{1}{\gamma \cdot \text{poly}(\log(1/\gamma))}\right).$$

In fact, we will prove a stronger result which holds also if  $C_A(\gamma)$  is not polynomial.

**Theorem 2.5** *Let  $A$  be a kernel-based learner for which  $C_A(\gamma) = \exp(o(\gamma^{-2/7}))$ . Then, for every  $\gamma > 0$ , there exists a distribution  $\mathcal{D}$  on  $B$  such that, w.p.  $\geq 1 - \exp(1/\gamma)$ ,*

$$\frac{\text{Err}_{\mathcal{D},0-1}(A(\gamma))}{\text{Err}_{\gamma}(\mathcal{D})} \geq \Omega\left(\frac{1}{\gamma \cdot \text{poly}(\log(C_A(\gamma)))}\right).$$

It is shown in (Birnbaum and Shalev-Shwartz, 2012) that solving kernel SVM with a specific kernel (i.e. a specific  $\psi$ ) yields an approximation ratio of  $O\left(\frac{1}{\gamma \sqrt{\log(1/\gamma)}}\right)$ . It follows that our lower bound in Theorem 2.5 is essentially tight. Also, this theorem can be viewed as a substantial generalization of (Ben-David et al., 2012, Long and Servedio, 2011), who

---

<sup>3</sup>Concretely, the needed sample size is  $O\left(\frac{(C \cdot L)^2 + 1/\rho}{\epsilon^2}\right)$  for  $L$  the Lipschitz constant of  $l$ . Since  $l$  is assumed to be predefined, we neglect the dependency on  $L$ .

give an approximation ratio of  $\Omega\left(\frac{1}{\gamma}\right)$  with no embedding (i.e.,  $\psi$  is the identity map). Also relevant is (Shalev-Shwartz et al., 2011), which shows that for a certain  $\psi$ , and  $C_A(\gamma) = \text{poly}(\exp((1/\gamma) \cdot \log(1/(\gamma))))$ , kernel SVM has approximation ratio of 1. Theorem 2.5 shows that, for some  $a > 0$ , for kernel SVM to achieve a constant approximation ratio,  $C_A$  must be  $\Omega(\exp((1/\gamma)^a))$ .

Next, we lower bound the approximation ratio of every efficient generalized linear learner.

**Theorem 2.6** *Let  $A$  be a generalized linear learner corresponding to a Lipschitz surrogate. Assume that  $m_A(\gamma) = \exp(o(\gamma^{-1/8}))$ . Then, for every  $\gamma > 0$ , there exists a distribution  $\mathcal{D}$  on  $S^{d-1} \times \{\pm 1\}$  with  $d = O(\log(m_A(\gamma)/\gamma))$  such that, w.p.  $\geq 1 - \exp(1/\gamma)$ ,*

$$\frac{\text{Err}_{\mathcal{D},0-1}(A(\gamma))}{\text{Err}_{\gamma}(\mathcal{D})} \geq \Omega\left(\frac{1}{\sqrt{\gamma} \text{poly}(\log(m_A(\gamma)/\gamma))}\right).$$

**Corollary 2.7** *Let  $A$  be an efficient generalized linear learner corresponding to a Lipschitz surrogate. Then, for every  $\gamma > 0$ , there exists a distribution  $\mathcal{D}$  on  $S^{d-1} \times \{\pm 1\}$  with  $d = O(\log(1/\gamma))$  such that, w.p.  $\geq 1 - \exp(1/\gamma)$ ,*

$$\frac{\text{Err}_{\mathcal{D},0-1}(A(\gamma))}{\text{Err}_{\gamma}(\mathcal{D})} \geq \Omega\left(\frac{1}{\sqrt{\gamma} \text{poly}(\log(1/\gamma))}\right).$$

## 2.2 Main proof ideas

To give the reader some idea of our arguments, we sketch some of the main ingredients of the proof of Theorem 2.5. At the end of this section we sketch the idea of the proof of Theorem 2.6. We note, however, that the actual proofs are organized somewhat differently.

We will construct a distribution  $\mathcal{D}$  over  $S^{d-1} \times \{\pm 1\}$  (recall that  $\mathbb{R}^d$  is viewed as standardly embedded in  $H = \ell^2$ ). Thus, we can assume that the program is formulated in terms of the unit sphere,  $S^\infty \subset \ell^2$ , and not the unit ball.

Fix an embedding  $\psi$  and  $C > 0$ . Denote by  $k : S^\infty \times S^\infty \rightarrow \mathbb{R}$  the corresponding kernel  $k(x, y) = \langle \psi(x), \psi(y) \rangle_{H_1}$  and consider the following set of functions over  $S^\infty$

$$H_k = \{\Lambda_{v,0} \circ \psi : v \in H_1\}.$$

$H_k$  is a Hilbert space with norm  $\|f\|_{H_k} = \inf\{\|v\|_{H_1} : \Lambda_{v,0} \circ \psi = f\}$ . The subscript  $k$  indicates that  $H_k$  is uniquely determined (as a Hilbert space) given the kernel  $k$ . With this interpretation, program (5) is equivalent to the program

$$\min_{f \in H_k, b \in \mathbb{R}} \text{Err}_{\mathcal{D},l}(f + b) \quad \text{s.t.} \quad \|f\|_{H_k} \leq C. \quad (6)$$

For simplicity we focus on  $l$  being the hinge-loss (the generalization to other surrogate loss functions is rather technical).

The proof may be split into three steps:



1. We consider the one-dimensional problem of improperly learning halfspaces (i.e. thresholds on the line) by optimizing the hinge loss over the space of univariate polynomials of degree bounded by  $\log(C)$ . We construct a distribution  $\mathcal{D}$  over  $[-1, 1] \times \{\pm 1\}$  that is a convex combination of two distributions. One that is separable by a  $\gamma$ -margin halfspace and the other representing a tiny amount of noise. We show that each solution of the problem of minimizing the hinge-loss w.r.t.  $\mathcal{D}$  over the space of such polynomials has the property that  $f(\gamma) \approx f(-\gamma)$ .
2. We pull back the distribution  $\mathcal{D}$  w.r.t. a direction  $e \in S^{d-1}$  to a distribution over  $S^{d-1} \times \{\pm 1\}$ . Let  $f$  be an approximate solution of program (6). We show that  $f$  takes almost the same value on instances for which  $\langle x, e \rangle = \gamma$  and  $\langle x, e \rangle = -\gamma$ . This step can be further broken into three substeps –
  - (a) First, we assume that the kernel is symmetric and  $f(x)$  depends only on  $\langle x, e \rangle$ . This substep uses a characterization of Hilbert spaces corresponding to symmetric kernels, from which it follows that  $f$  has the form

$$f(x) = \sum_{n=1}^{\infty} \alpha_n P_{d,n}(\langle x, e \rangle) .$$

Here  $P_{d,n}$  are the  $d$ -dimensional Legendre polynomials and  $\sum_{n=0}^{\infty} \alpha_n^2 < C^2$ . This allows us to rely on the results for the one-dimensional case from step (1).

- (b) By symmetrizing  $f$ , we relax the assumption that  $f$  depends only on  $\langle x, e \rangle$ .
  - (c) By averaging the kernel over the group of linear isometries on  $\mathbb{R}^d$ , we relax the assumption that the kernel is symmetric.
3. Finally, we show that for the distribution from the previous step, if  $f$  is an approximate solution to program (6) then  $f$  predicts the same value, 1, on instances for which  $\langle x, e \rangle = \gamma$  and  $\langle x, e \rangle = -\gamma$ . This establishes our claim, as the constructed distribution assigns the value  $-1$  to instances for which  $\langle x, e \rangle = -\gamma$ .

We now expand on this brief description of the main steps.

### The one dimensional distribution

We define a distribution  $\mathcal{D}$  on  $[-1, 1]$  as follows. Start with the distribution  $\mathcal{D}_1$  that takes the values  $\pm(\gamma, 1)$ , where  $\mathcal{D}_1(\gamma, 1) = 0.7$  and  $\mathcal{D}_1(-\gamma, -1) = 0.3$ . Clearly, for this distribution, the threshold 0 has zero error rate. To construct  $\mathcal{D}$ , we perturb  $\mathcal{D}_1$  with “noise” as follows. Let  $\mathcal{D} = (1 - \lambda)\mathcal{D}_1 + \lambda\mathcal{D}_2$ , where  $\mathcal{D}_2$  is defined as follows. The probability of the labels is uniform and independent of the instance and the marginal probability over the instances is defined by the density function

$$\rho(x) = \begin{cases} 0 & \text{if } |x| > 1/8 \\ \frac{8}{\pi\sqrt{1-(8x)^2}} & \text{if } |x| \leq 1/8 \end{cases} .$$



This choice of  $\rho$  simplifies our calculations due to its relation to Chebyshev polynomials. However, other choices of  $\rho$  which are supported on a small interval around zero can also work.

Note that the error rate of the threshold 0 on  $\mathcal{D}$  is  $\lambda/2$ . We next show that each polynomial  $f$  of degree  $K = \log(C)$  that satisfies  $\text{Err}_{\mathcal{D}, \text{hinge}}(f) \leq 1$  must have  $f(\gamma) \approx f(-\gamma)$ . Indeed, if

$$1 \geq \text{Err}_{\mathcal{D}, \text{hinge}}(f) = (1 - \lambda) \text{Err}_{\mathcal{D}_1, \text{hinge}}(f) + \lambda \text{Err}_{\mathcal{D}_2, \text{hinge}}(f)$$

then  $\text{Err}_{\mathcal{D}_2, \text{hinge}}(f) \leq \frac{1}{\lambda}$ . But,

$$\begin{aligned} \text{Err}_{\mathcal{D}_2, \text{hinge}}(f) &= \frac{1}{2} \int_{-1}^1 l_{\text{hinge}}(f(x)) \rho(x) dx + \frac{1}{2} \int_{-1}^1 l_{\text{hinge}}(-f(x)) \rho(x) dx \\ &\geq \frac{1}{2} \int_{-1}^1 l_{\text{hinge}}(-|f(x)|) \rho(x) dx \end{aligned}$$

and using the convexity of  $l_{\text{hinge}}$  we obtain from Jensen's inequality that

$$\begin{aligned} &\geq \frac{1}{2} l_{\text{hinge}} \left( \int_{-1}^1 -|f(x)| \rho(x) dx \right) \\ &= \frac{1}{2} \left( 1 + \int_{-1}^1 |f(x)| \rho(x) dx \right) \\ &\geq \frac{1}{2} \int_{-1}^1 |f(x)| \rho(x) dx =: \frac{1}{2} \|f\|_{1, d\rho} . \end{aligned}$$

This shows that  $\|f\|_{1, d\rho} \leq \frac{2}{\lambda}$ . We next write  $f = \sum_{i=1}^K \alpha_i \tilde{T}_i$ , where  $\{\tilde{T}_i\}$  are the orthonormal polynomials corresponding to the measure  $d\rho$ . Since  $\tilde{T}_i$  are related to Chebyshev polynomials we can uniformly bound their  $\ell_\infty$  norm, hence obtain that

$$\sqrt{\sum_i \alpha_i^2} = \|f\|_{2, d\rho} \leq O(\sqrt{k}) \|f\|_{1, d\rho} \leq O\left(\frac{\sqrt{K}}{\lambda}\right) .$$

Based on the above, and using a bound on the derivatives of Chebyshev polynomials, we can bound the derivative of the polynomial  $f$

$$|f'(x)| \leq \sum_i |\alpha_i| |\tilde{T}_i'(x)| \leq O\left(\frac{K^3}{\lambda}\right) .$$

Hence, by choosing  $\lambda = \omega(\gamma K^3) = \omega(\gamma \log^3(C))$  we obtain

$$|f(\gamma) - f(-\gamma)| \leq 2\gamma \max_x |f'(x)| = O\left(\frac{\gamma K^3}{\lambda}\right) = o(1) ,$$

as required.

## Pulling back to the $d - 1$ dimensional sphere

Given the distribution  $\mathcal{D}$  over  $[-1, 1] \times \{\pm 1\}$  described before, and some  $e \in S^{d-1}$ , we now define a distribution  $\mathcal{D}_e$  on  $S^{d-1} \times \{\pm 1\}$ . To sample from  $\mathcal{D}_e$ , we first sample  $(\alpha, \beta)$  from  $\mathcal{D}$  and (uniformly and independently) a vector  $z$  from the 1-codimensional sphere of  $S^{d-1}$  that is orthogonal to  $e$ . The constructed point is  $(\alpha e + \sqrt{1 - \alpha^2}z, \beta)$ .

For any  $f \in H_k$  and  $a \in [-1, 1]$  define  $\bar{f}(a)$  to be the expectation of  $f$  over the 1-codimensional sphere  $\{x \in S^{d-1} : \langle x, e \rangle = a\}$ . We will show that for any  $f \in H_k$ , such that  $\|f\|_{H_k} \leq C$  and  $\text{Err}_{\mathcal{D}_e, \text{hinge}}(f) \leq 1$ , we have that  $|\bar{f}(\gamma) - \bar{f}(-\gamma)| = o(1)$ .

To do so, let us first assume that  $f$  is symmetric with respect to  $e$ , and hence can be written as

$$f(x) = \sum_{n=0}^{\infty} \alpha_n P_{d,n}(\langle x, e \rangle) ,$$

where  $\alpha_n \in \mathbb{R}$  and  $P_{d,n}$  is the  $d$ -dimensional Legendre polynomial of degree  $n$ . Furthermore, by a characterization of Hilbert spaces corresponding to symmetric kernels, it follows that  $\sum \alpha_n^2 \leq C^2$ .

Since  $f$  is symmetric w.r.t.  $e$  we have,

$$\bar{f}(a) = \sum_{n=0}^{\infty} \alpha_n P_{d,n}(a) .$$

For  $|a| \leq 1/8$ , we have that  $|P_{d,n}(a)|$  tends to zero exponentially fast with both  $d$  and  $n$ . Hence, if  $d$  is large enough then

$$\bar{f}(a) \approx \sum_{n=0}^{\log(C)} \alpha_n P_{d,n}(a) =: \tilde{f}(a) .$$

Note that  $\tilde{f}$  is a polynomial of degree bounded by  $\log(C)$ . In addition, by construction,  $\text{Err}_{\mathcal{D}_e, \text{hinge}}(f) = \text{Err}_{\mathcal{D}, \text{hinge}}(\bar{f}) \approx \text{Err}_{\mathcal{D}, \text{hinge}}(\tilde{f})$ . Hence, if  $1 \geq \text{Err}_{\mathcal{D}_e, \text{hinge}}(f)$  then using the previous subsection we conclude that  $|\bar{f}(\gamma) - \bar{f}(-\gamma)| = o(1)$ .

## Symmetrization of $f$

In the above, we assumed that both the kernel function is symmetric and that  $f$  is symmetric w.r.t.  $e$ . Our next step is to relax the latter assumption, while still assuming that the kernel function is symmetric.

Let  $\mathbb{O}(e)$  be the group of linear isometries that fix  $e$ , namely,  $\mathbb{O}(e) = \{A \in \mathbb{O}(d) : Ae = e\}$ . By assuming that  $k$  is a symmetric kernel, we have that for all  $A \in \mathbb{O}(e)$ , the function  $g(x) = f(Ax)$  is also in  $H_k$ . Furthermore,  $\|g\|_{H_k} = \|f\|_{H_k}$  and by the construction of  $\mathcal{D}_e$  we also have  $\text{Err}_{\mathcal{D}_e, \text{hinge}}(g) = \text{Err}_{\mathcal{D}_e, \text{hinge}}(f)$ . Let  $\mathcal{P}_e f(x) = \int_{\mathbb{O}(e)} f(Ax) dA$  be the symmetrization of  $f$  w.r.t.  $e$ . On one hand,  $\mathcal{P}_e f \in H_k$ ,  $\|\mathcal{P}_e f\|_{H_k} \leq \|f\|_{H_k}$ , and  $\text{Err}_{\mathcal{D}_e, \text{hinge}}(\mathcal{P}_e f) \leq \text{Err}_{\mathcal{D}_e, \text{hinge}}(f)$ . On the other hand,  $\bar{f} = \overline{\mathcal{P}_e f}$ . Since for  $\mathcal{P}_e f$  we have already shown that  $|\overline{\mathcal{P}_e f}(\gamma) - \overline{\mathcal{P}_e f}(-\gamma)| = o(1)$ , it follows that  $|\bar{f}(\gamma) - \bar{f}(-\gamma)| = o(1)$  as well.

## Symmetrization of the kernel

Our final step is to remove the assumption that the kernel is symmetric. To do so, we first symmetrize the kernel as follows. Recall that  $\mathbb{O}(d)$  is the group of linear isometries of  $\mathbb{R}^d$ . Define the following symmetric kernel:

$$k_s(x, y) = \int_{\mathbb{O}(d)} k(Ax, Ay) dA .$$

We show that the corresponding Hilbert space consists of functions of the form

$$f(x) = \int_{\mathbb{O}(d)} f_A(Ax) dA ,$$

where for every  $A$   $f_A \in H_k$ . Moreover,

$$\|f\|_{H_{k_s}}^2 \leq \int_{\mathbb{O}(d)} \|f_A\|_{H_k}^2 dA . \quad (7)$$

Let  $\alpha$  be the maximal number such that

$$\forall e \in S^{d-1} \exists f_e \in H_k \text{ s.t. } \|f_e\|_{H_k} \leq C, \text{ Err}_{D_e, \text{hinge}}(f_e) \leq 1, |\bar{f}_e(\gamma) - \bar{f}_e(-\gamma)| > \alpha .$$

Since  $H_k$  is closed to negation, it follows that  $\alpha$  satisfies

$$\forall e \in S^{d-1} \exists f_e \in H_k \text{ s.t. } \|f_e\|_{H_k} \leq C, \text{ Err}_{D_e, \text{hinge}}(f_e) \leq 1, \bar{f}_e(\gamma) - \bar{f}_e(-\gamma) > \alpha .$$

Fix some  $v \in S^{d-1}$  and define  $f \in H_{k_s}$  to be

$$f(x) = \int_{\mathbb{O}(d)} f_{Av}(Ax) dA .$$

By Equation (7) we have that  $\|f\|_{H_{k_s}} \leq C$ . It is also possible to show that for all  $A$   $\text{Err}_{\mathcal{D}_v, \text{hinge}}(f_{Av} \circ A) = \text{Err}_{\mathcal{D}_{Av}, \text{hinge}}(f_{Av}) \leq 1$ . Therefore, by the convexity of the loss,  $\text{Err}_{\mathcal{D}_v, \text{hinge}}(f) \leq 1$ . It follows, by the previous sections, that  $|\bar{f}(\gamma) - \bar{f}(-\gamma)| = o(1)$ . But, we show that  $\bar{f}(\gamma) - \bar{f}(-\gamma) > \alpha$ . It therefore follows that  $\alpha = o(1)$ , as required.

## Concluding the proof

We have shown that for every kernel, there exists some direction  $e$  such that for all  $f \in H_k$  that satisfies  $\|f\|_{H_k} \leq C$  and  $\text{Err}_{\mathcal{D}_e, \text{hinge}}(f) \leq 1$  we have that  $|\bar{f}(\gamma) - \bar{f}(-\gamma)| = o(1)$ .

Next, consider  $f$  which is also an (approximated) optimal solution of program (2) with respect to  $\mathcal{D}_e$ . Since  $\text{Err}_{\mathcal{D}_e, \text{hinge}}(0) = 1$ , we clearly have that  $\text{Err}_{\mathcal{D}_e, \text{hinge}}(f) \leq 1$ , hence  $|\bar{f}(\gamma) - \bar{f}(-\gamma)| = o(1)$ . Next we show that  $\bar{f}(-\gamma) > 1/2$ , which will imply that  $f$  predicts the label 1 for most instances on the 1 co-dimensional sphere such that  $\langle x, e \rangle = -\gamma$ . Hence, its 0-1 error is close to  $0.3(1-\lambda) \geq 0.2$  while  $\text{Err}_\gamma(\mathcal{D}_e) = \lambda/2$ . By choosing  $\lambda = O(\gamma \log^{3.1}(C))$  we obtain that the approximation ratio is  $\Omega\left(\frac{1}{\gamma \log^{3.1}(C)}\right)$ .

It is therefore left to show that  $\bar{f}(-\gamma) > 1/2$ . Let  $a = \bar{f}(\gamma) \approx \bar{f}(-\gamma)$ . On  $(1-\lambda)$  fraction of the distribution, the hinge-loss would be (on average and roughly)  $0.3[1+a]_+ + 0.7[1-a]_+$ . This function is minimized for  $a = 1$ , which concludes our proof since  $\lambda$  is  $o(1)$ .

## The proof of Theorem 2.6

To prove Theorem 2.6, we prove, using John's Lemma (Matousek, 2002), that for every embedding  $\psi : S^{d-1} \rightarrow B_1$ , we can construct a kernel  $k : S^{d-1} \times S^{d-1} \rightarrow \mathbb{R}$  and a probability measure  $\mu_N$  over  $S^{d-1}$  with the following properties: If  $f$  is an approximate solution of program (4), where  $\gamma$  fraction of the distribution  $\mathcal{D}$  is perturbed by  $\mu_N$ , then  $\|f\|_k \leq O\left(\frac{m^{1.5}}{\gamma}\right)$ . Using this, we adapt the proof as sketched above to prove Theorem 2.6.

## 3 Additional Results

**Low dimensional distributions.** It is of interest to examine Theorem 2.5 when  $\mathcal{D}$  is supported on  $B^d$  for  $d$  small. We show that for  $d = O(\log(1/\gamma))$ , the approximation ratio is  $\Omega\left(\frac{1}{\sqrt{\gamma} \cdot \text{poly}(\log(1/\gamma))}\right)$ . Most commonly used kernels (e.g., the polynomial, RBF, and Hyperbolic tangent kernels, as well as the kernel used in (Shalev-Shwartz et al., 2011)) are symmetric. Namely, for all unit vectors  $x, y \in B$ ,  $k(x, y) := \langle \psi(x), \psi(y) \rangle_{H_1}$  depends only on  $\langle x, y \rangle_H$ . For symmetric kernels, we show that even with the restriction that  $d = O(\log(1/\gamma))$ , the approximation ratio is still  $\Omega\left(\frac{1}{\gamma \cdot \text{poly}(\log(1/\gamma))}\right)$ . However, the result for symmetric kernels is only proved for (idealized) algorithms that return the exact solution of program (5).

**Theorem 3.1** *Let  $A$  be a kernel-based learner corresponding to a Lipschitz surrogate. Assume that  $C_A(\gamma) = \exp(o(\gamma^{-1/8}))$ . Then, for every  $\gamma > 0$ , there exists a distribution  $\mathcal{D}$  on  $B^d$ , for  $d = O(\log(C_A(\gamma)/\gamma))$ , such that, w.p.  $\geq 1 - \exp(-1/\gamma)$ ,*

$$\frac{\text{Err}_{\mathcal{D},0-1}(A(\gamma))}{\text{Err}_{\gamma}(\mathcal{D})} \geq \Omega\left(\frac{1}{\sqrt{\gamma} \cdot \text{poly}(\log(C_A(\gamma)))}\right).$$

**Theorem 3.2** *Assume that  $C = \exp(o(\gamma^{-2/7}))$  and  $\psi$  is continuous and symmetric. For every  $\gamma > 0$ , there exists a distribution  $\mathcal{D}$  on  $B^d$ , for  $d = O(\log(C))$  and a solution to program (5) whose 0-1-error is  $\Omega\left(\frac{1}{\gamma \text{poly}(\log(C))}\right) \cdot \text{Err}_{\gamma}(\mathcal{D})$ .*

**The integrality gap.** In bounding the approximation ratio, we considered a predefined loss  $l$ . We believe that similar bounds hold as well for algorithms that can choose  $l$  according to  $\gamma$ . However, at the moment, we only know to lower bound the *integrality gap*, as defined below.

If we let  $l$  depend on  $\gamma$ , we should redefine the complexity of Program (5) to be  $C \cdot L$ , where  $L$  is the Lipschitz constant of  $l$ . (See the discussion following Program (5)). The  $(\gamma)$ -*integrality gap* of program (5) and (4) is defined as the worst case, over all possible choices of  $\mathcal{D}$ , of the ratio between the optimum of the program, running on the input  $\gamma$ , to  $\text{Err}_{\gamma}(\mathcal{D})$ . We note that  $\text{Err}_{\mathcal{D},0-1}(f) \leq \text{Err}_{\mathcal{D},l}(f)$  for every convex surrogate  $l$ . Thus, the integrality gap always upper bounds the approximation ratio. Moreover, this fact establishes most (if not all) guarantees for algorithms that solve Program (5) or Program (4).

We denote by  $\partial_+ f$  the right derivative of the real function  $f$ . Note that  $\partial_+ f$  always exists for  $f$  convex. Also,  $\forall x \in \mathbb{R}$ ,  $|\partial_+ f(x)| \leq L$  if  $f$  is  $L$ -Lipschitz. We prove:

**Theorem 3.3** Assume that  $C = \exp(o(\gamma^{-2/7}))$  and  $\psi$  is continuous. For every  $\gamma > 0$ , there exists a distribution  $\mathcal{D}$  on  $B^d$ , for  $d = O(\log(C))$  such that the optimum of Program (5) is  $\Omega\left(\frac{1}{\gamma \text{poly}(\log(C \cdot |\partial_+ l(0)|))}\right) \cdot \text{Err}_\gamma(\mathcal{D})$ .

Thus Program (5) has integrality gap  $\Omega\left(\frac{1}{\gamma \text{poly}(\log(C \cdot L))}\right)$ . For Program (4) we prove a similar lower bound:

**Theorem 3.4** Let  $m, d, \gamma$  such that  $d = \omega(\log(m/\gamma))$  and  $m = \exp(o(\gamma^{-2/7}))$ . There exist a distribution  $\mathcal{D}$  on  $S^{d-1} \times \{\pm 1\}$  such that the optimum of Program (4) is  $\Omega\left(\frac{1}{\gamma \text{poly}(\log(m/\gamma))}\right) \cdot \text{Err}_\gamma(\mathcal{D})$ .

## 4 Conclusion

We gave lower bounds on the worst-case approximation ratio of the family of methods in which one minimizes a convex surrogate loss over affine functionals in some feature space. This applies to many popular algorithms such as kernel SVM, with any kernel, and the Lasso, with any feature mapping. Furthermore, our lower bound nearly matches the best known upper bound. Our analysis suggests that if better approximation ratios are at all possible, they would require methods other than optimizing a convex surrogate over a regularized linear class of classifiers.

The main spirit of our results is that from a worst case perspective, Kernel-SVM is a rather bad algorithm. Nevertheless, it is an empirical fact that Kernel-SVM performs very well in practice. Therefore, it is of great interest to find distributional conditions that hold in practice, under which Kernel-SVM is guaranteed to perform well. We note that learning halfspaces under distributional assumptions, has already been addressed. For example, (Kalai et al., 2005, Blais et al., 2008) show positive results under several assumptions on the marginal distribution. However, to the best of our knowledge, these results yield runtime which is exponential in  $\text{poly}(1/\epsilon)$ , where  $\epsilon$  is the excess error of the learnt hypothesis.

**Open questions:** Our analysis is limited in several ways. First, we assumed that the surrogate loss is fixed in advance, but we believe that results similar to ours hold even if the loss may depend on  $\gamma$ . This belief is supported by our results about the integrality gap. As explained in Section 6, this is somewhat subtle issue and related to questions about sample complexity. Also, we refer to  $C$  as the complexity of Program (3) since the analysis of uniform convergence requires a sample of size  $\text{poly}(C)$  in order to solve the problem based on a finite sample. Our results do not rule out the possibility of choosing  $C$  that is exponentially large in  $1/\gamma$  and still using a polynomial sample. We believe that this approach is doomed to fail due to over-fitting. Third, in view of Theorems 3.3 and 3.4, we believe that the lower bound in Theorems 2.6 and 3.1 can be improved to depend on  $\frac{1}{\gamma}$  rather than on  $\frac{1}{\sqrt{\gamma}}$ , as in Theorem 2.5.

## 5 Proofs

### 5.1 Background and Notation

Here we introduce some notations and terminology to be used throughout. The  $L^p$  norm corresponding to a measure  $\mu$  is denoted  $\|\cdot\|_{p,\mu}$ . Also,  $\mathbb{N} = \{1, 2, \dots\}$  and  $\mathbb{N}_0 = \{0, 1, 2, \dots\}$ .

#### 5.1.1 Reproducing Kernel Hilbert Spaces

All the theorems we quote here are standard and can be found, e.g., in Chapter 2 of (Saitoh, 1988). Let  $H$  be a Hilbert space of functions from a set  $S$  to  $\mathbb{C}$ . Note that  $H$  consists of functions and not of equivalence classes of functions. We say that  $H$  is a *reproducing kernel Hilbert space* (RKHS for short) if, for every  $x \in S$ , the linear functional  $f \rightarrow f(x)$  is bounded.

A function  $k : S \times S \rightarrow \mathbb{C}$  is a *reproducing kernel* (or just a *kernel*) if, for every  $x_1, \dots, x_n \in S$ , the matrix  $\{k(x_i, x_j)\}_{1 \leq i, j \leq n}$  is positive semi-definite.

Kernels and RKHSs are essentially synonymous:

**Theorem 5.1** 1. For every kernel  $k$  there exists a unique RKHS  $H_k$  such that for every  $y \in S$ ,  $k(\cdot, y) \in H_k$  and  $\forall f \in H$ ,  $f(y) = \langle f(\cdot), k(\cdot, y) \rangle_{H_k}$ .

2. A Hilbert space  $H \subseteq \mathbb{C}^S$  is a RKHS if and only if there exists a kernel  $k : S \times S \rightarrow \mathbb{R}$  such that  $H = H_k$ .

3. For every kernel  $k$ ,  $\overline{\text{span}\{k(\cdot, y)\}_{y \in S}} = H_k$ . Moreover,

$$\left\langle \sum_{i=1}^n \alpha_i k(\cdot, x_i), \sum_{i=1}^n \beta_i k(\cdot, y_i) \right\rangle_{H_k} = \sum_{1 \leq i, j \leq n} \alpha_i \bar{\beta}_j k(y_j, x_i)$$

4. If the kernel  $k : S \times S \rightarrow \mathbb{R}$  takes only real values, then  $H_k^{\mathbb{R}} := \{\text{Re}(f) : f \in H_k\} \subset H_k$ . Moreover,  $H_k^{\mathbb{R}}$  is a real Hilbert space with the inner product induced from  $H_k$ .

5. For every kernel  $k$ , convergence in  $H_k$  implies point-wise convergence. If  $\sup_{x \in S} k(x, x) < \infty$  then this convergence is uniform.

There is also a tight connection between embeddings of  $S$  into a Hilbert space and RKHSs.

**Theorem 5.2** A function  $k : S \times S \rightarrow \mathbb{R}$  is a kernel iff there exists a mapping  $\phi : S \rightarrow H$  to some real Hilbert space for which  $k(x, y) = \langle \phi(y), \phi(x) \rangle_H$ . Also,

$$H_k = \{f_v : v \in H\}$$

Where  $f_v(x) = \langle v, \phi(x) \rangle_H$ . The mapping  $v \mapsto f_v$ , restricted to  $\overline{\text{span}(\phi(S))}$ , is a Hilbert space isomorphism.

A kernel  $k : S \times S \rightarrow \mathbb{R}$  is called *normalized* if  $\sup_{x \in S} k(x, x) = 1$ . Also,

**Theorem 5.3** Let  $k : S \times S \rightarrow \mathbb{R}$  be a kernel and let  $\{f_n\}_{n=1}^{\infty}$  be an orthonormal basis of a  $H_k$ . Then,  $k(x, y) = \sum_{n=1}^{\infty} f_n(x) f_n(y)$ .

### 5.1.2 Unitary Representations of Compact Groups

Proofs of the results stated here can be found in (Folland, 1994), chapter 5. Let  $G$  be a compact group. A *unitary representation* (or just a *representation*) of  $G$  is a group homomorphism  $\rho : G \rightarrow U(H)$  where  $U(H)$  is the class of unitary operators over a Hilbert space  $H$ , such that, for every  $v \in H$ , the mapping  $a \mapsto \rho(a)v$  is continuous.

We say that a closed subspace  $M \subset H$  is *invariant* (to  $\rho$ ) if for every  $a \in G, v \in M$ ,  $\rho(a)v \in M$ . We note that if  $M$  is invariant then so is  $M^\perp$ . We denote by  $\rho|_M : G \rightarrow U(M)$  the *restriction* of  $\rho$  to  $M$ . That is,  $\forall a \in G, \rho|_M(a) = \rho(a)|_M$ . We say that  $\rho : G \rightarrow U(H)$  is *reducible* if  $H = M \oplus M^\perp$  such that  $M, M^\perp$  are both non zero closed and invariant subspaces of  $H$ . A basic result is that every representation of a compact group is a sum of irreducible representation.

**Theorem 5.4** *Let  $\rho : G \rightarrow U(H)$  be a representation of a compact group  $G$ . Then,  $H = \bigoplus_{n \in I} H_n$ , where every  $H_n$  is invariant to  $\rho$  and  $\rho|_{H_n}$  is irreducible.*

We shall also use the following Lemma.

**Lemma 5.5** *Let  $G$  be a compact group,  $V$  a finite dimensional vector space and let  $\rho : G \rightarrow GL(V)$  be a continuous homomorphism of groups (here,  $GL(V)$  is the group of invertible linear operators over  $V$ ). Then,*

1. *There exists an inner product on  $V$  making  $\rho$  a unitary representation.*
2. *Moreover, if  $V$  has no non-trivial invariant subspaces (here a subspace  $U \subset V$  is called invariant if,  $\forall a \in G, f \in U, \rho(a)f \in U$ ) then this inner product is unique up to scalar multiple.*

### 5.1.3 Harmonic Analysis on the Sphere

All the results stated here can be found in (Atkinson and Han, 2012), chapters 1 and 2. Denote by  $\mathbb{O}(d)$  the group of unitary operators over  $\mathbb{R}^d$  and by  $dA$  the uniform probability measure over  $\mathbb{O}(d)$  (that is,  $dA$  is the unique probability measure satisfying  $\int_{\mathbb{O}(d)} f(A)dA = \int_{\mathbb{O}(d)} f(AB)dA = \int_{\mathbb{O}(d)} f(BA)dA$  for every  $B \in \mathbb{O}(d)$  and every integrable function  $f : \mathbb{O}(d) \rightarrow \mathbb{C}$ ). Denote by  $dx = dx_{d-1}$  the Lebesgue (area) measure over  $S^{d-1}$  and let  $L^2(S^{d-1}) := L^2(S^{d-1}, dx)$ . Given a measurable set  $Z \subseteq S^{d-1}$ , we sometime denote its Lebesgue measure by  $|Z|$ . Also, denote  $dm = \frac{dx}{|S^{d-1}|}$  the Lebesgue measure, normalized to be a probability measure.

For every  $n \in \mathbb{N}_0$ , we denote by  $\mathbb{Y}_n^d$  the linear space of  $d$ -variables harmonic (i.e., satisfying  $\Delta p = 0$ ) homogeneous polynomials of degree  $n$ . It holds that

$$N_{d,n} = \dim(\mathbb{Y}_n^d) = \binom{d+n-1}{d-1} - \binom{d+n-3}{d-1} = \frac{(2n+d-2)(n+d-3)!}{n!(d-2)!} \quad (8)$$

Denote by  $\mathcal{P}_{d,n} : L^2(S^{d-1}) \rightarrow \mathbb{Y}_n^d$  the orthogonal projection onto  $\mathbb{Y}_n^d$ .

We denote by  $\rho : \mathbb{O}(d) \rightarrow U(L^2(S^{d-1}))$  the unitary representation defined by

$$\rho(A)f = f \circ A^{-1}$$



We say that a closed subspace  $M \subset L^2(S^{d-1})$  is *invariant* if it is invariant w.r.t.  $\rho$  (that is,  $\forall f \in M, A \in \mathbb{O}(d), f \circ A \in M$ ). We say that an invariant space  $M$  is *primitive* if  $\rho|_M$  is irreducible.

**Theorem 5.6** 1.  $L^2(S^{d-1}) = \oplus_{n=0}^{\infty} \mathbb{Y}_n^d$ .

2. The primitive finite dimensional subspaces of  $L^2(S^{d-1})$  are exactly  $\{\mathbb{Y}_n^d\}_{n=0}^{\infty}$ .

**Lemma 5.7** Fix an orthonormal basis  $Y_{n,j}^d$ ,  $1 \leq j \leq N_{d,n}$  to  $\mathbb{Y}_n^d$ . For every  $x \in S^{d-1}$  it holds that

$$\sum_{j=1}^{N_{d,n}} |Y_{n,j}^d(x)|^2 = \frac{N_{d,n}}{|S^{d-1}|}$$

#### 5.1.4 Legendre and Chebyshev Polynomials

The results stated here can be found at (Atkinson and Han, 2012). Fix  $d \geq 2$ . The  $d$  dimensional Legendre polynomials are the sequence of polynomials over  $[-1, 1]$  defined by the recursion formula

$$P_{d,n}(x) = \frac{2n+d-4}{n+d-3} x P_{d,n-1}(x) + \frac{n-1}{n+d-3} P_{d,n-2}(x)$$

$$P_{d,0} \equiv 1, P_{d,1}(x) = x$$

We shall make use of the following properties of the Legendre polynomials.

**Proposition 5.8** 1. For every  $d \geq 2$ , the sequence  $\{P_{d,n}\}$  is orthogonal basis of the Hilbert space  $L^2\left([-1, 1], (1-x^2)^{\frac{d-3}{2}} dx\right)$ .

2. For every  $n, d$ ,  $\|P_{d,n}\|_{\infty} = 1$ .

The Chebyshev polynomials of the first kind are defined as  $T_n := P_{2,n}$ . The Chebyshev polynomials of the second kind are the polynomials over  $[-1, 1]$  defined by the recursion formula

$$U_n(x) = 2xU_{n-1}(x) - U_{n-2}(x)$$

$$U_0 \equiv 1, U_1(x) = 2x$$

We shall make use of the following properties of the Chebyshev polynomials.

**Proposition 5.9** 1. For every  $n \geq 1$ ,  $T'_n = nU_{n-1}$ .

2.  $\|U_n\|_{\infty} = n + 1$ .

Given a measure  $\mu$  over  $[-1, 1]$ , the orthogonal polynomials corresponding to  $\mu$  are the sequence of polynomials obtained upon the Gram-Schmidt procedure applied to  $1, x, x^2, x^3, \dots$ . We note that the  $1, \sqrt{2}T_1, \sqrt{2}T_2, \sqrt{2}T_3, \dots$  are the orthogonal polynomials corresponding to the probability measure  $d\mu = \frac{dx}{\pi\sqrt{1-x^2}}$

### 5.1.5 Bochner Integral and Bochner Spaces

Proofs and elaborations on the material appearing in this section can be found in (Kosaku Yosida, 1963). Let  $(X, \mathfrak{m}, \mu)$  be a measure space and let  $H$  be a Hilbert space. A function  $f : X \rightarrow H$  is (Bochner) measurable if there exists a sequence of function  $f_n : X \rightarrow H$  such that

- For almost every  $x \in X$ ,  $f(x) = \lim_{n \rightarrow \infty} f_n(x)$ .
- The range of every  $f_n$  is countable and, for every  $v \in H$ ,  $f^{-1}(v)$  is measurable.

A measurable function  $f : X \rightarrow H$  is (Bochner) integrable if there exists a sequence of simple measurable functions (in the usual sense)  $s_n$  such that  $\lim_{n \rightarrow \infty} \int_X \|f(x) - s_n(x)\|_H d\mu(x) = 0$ . We define the integral of  $f$  to be  $\int_X f d\mu = \lim_{n \rightarrow \infty} \int s_n d\mu$ , where the integral of a simple function  $s = \sum_{i=1}^n 1_{A_i} v_i$ ,  $A_i \in \mathfrak{m}$ ,  $v_i \in H$  is  $\int_X s d\mu = \sum_{i=1}^n \mu(A_i) v_i$ .

Define by  $L^2(X, H)$  the Kolmogorov quotient (by equality almost everywhere) of all measurable functions  $f : X \rightarrow H$  such that  $\int_X \|f\|_H^2 d\mu < \infty$ .

**Theorem 5.10**  $L^2(X, H)$  in a Hilbert space w.r.t. the inner product  $\langle f, g \rangle_{L^2(X, H)} = \int_X \langle f(x), g(x) \rangle_H d\mu(x)$

## 5.2 Symmetric Kernels and Symmetrization

In this section we concern *symmetric kernels*. Fix  $d \geq 2$  and let  $k : S^{d-1} \times S^{d-1} \rightarrow \mathbb{R}$  be a continuous positive definite kernel. We say that  $k$  is *symmetric* if

$$\forall A \in \mathbb{O}(d), x, y \in S^{d-1}, k(Ax, Ay) = k(x, y)$$

In other words,  $k(x, y)$  depends only on  $\langle x, y \rangle_{\mathbb{R}^d}$ . A RKHS is called *symmetric* if its reproducing kernel is symmetric. The next theorem characterizes symmetric RKHSs. We note that Theorems of the same spirit have already been proved (e.g. (Schoenberg, 1942)).

**Theorem 5.11** Let  $k : S^{d-1} \times S^{d-1} \rightarrow \mathbb{R}$  be a normalized, symmetric and continuous kernel. Then,

1. The group  $\mathbb{O}(d)$  acts on  $H_k$ . That is, for every  $A \in \mathbb{O}(d)$  and every  $f \in H_k$  it holds that  $f \circ A \in H_k$  and  $\|f\|_{H_k} = \|f \circ A\|_{H_k}$ .
2. The mapping  $\rho : \mathbb{O}(d) \rightarrow U(H_k)$  defined by  $\rho(A)f = f \circ A^{-1}$  is a unitary representation.
3. The space  $H_k$  consists of continuous functions.
4. The decomposition of  $\rho$  into a sum of irreducible representation is  $H = \bigoplus_{n \in I} \mathbb{Y}_n^d$  for some set  $I \subset \mathbb{N}_0$ . Moreover,

$$\forall f, g \in H_k, \langle f, g \rangle_{H_k} = \sum_{n \in I} a_n^2 \langle \mathcal{P}_{d,n} f, \mathcal{P}_{d,n} g \rangle_{L^2(S^{d-1})}$$

Where  $\{a_n\}_{n \in I}$  are positive numbers.

5. It holds that  $\sum_{n \in I} \frac{N_{d,n}}{|S^{d-1}|} a_n^{-2} = 1$ .

**Proof** Let  $f \in H_k$ ,  $A \in \mathbb{O}(d)$ . To prove part 1, assume first that

$$\forall x \in S^{d-1}, f(x) = \sum_{i=1}^n \alpha_i k(x, y_i) \quad (9)$$

For some  $y_1, \dots, y_n \in S^{d-1}$  and  $\alpha_1, \dots, \alpha_n \in \mathbb{C}$ . We have, since  $k$  is symmetric, that

$$\begin{aligned} f \circ A(x) &= \sum_{i=1}^n \alpha_i k(Ax, y_i) \\ &= \sum_{i=1}^n \alpha_i k(A^{-1}Ax, A^{-1}y_i) \\ &= \sum_{i=1}^n \alpha_i k(x, A^{-1}y_i) \end{aligned}$$

Thus, by Theorem 5.1,  $f \circ A \in H_k$ . Moreover, it holds that

$$\begin{aligned} \|f \circ A\|_{H_k}^2 &= \sum_{1 \leq i, j \leq n} \alpha_i \bar{\alpha}_j k(A^{-1}y_j, A^{-1}y_i) \\ &= \sum_{1 \leq i, j \leq n} \alpha_i \bar{\alpha}_j k(y_j, y_i) = \|f\|_{H_k}^2 \end{aligned}$$

Thus, part 1 holds for function  $f \in H_k$  of the form (9). For general  $f \in H_k$ , by Theorem 5.1, there is a sequence  $f_n \in H_k$  of functions of the form (9) that converges to  $f$  in  $H_k$ . From what we have shown for functions of the form (9) it follows that  $\|f_n - f_m\|_{H_k} = \|f_n \circ A - f_m \circ A\|_{H_k}$ , thus  $f_n \circ A$  is a Cauchy sequence, hence, has a limit  $g \in H_k$ . By Theorem 5.1, convergence in  $H_k$  entails point wise convergence, thus,  $g = f \circ A$ . Finally,

$$\|f\|_{H_k} = \lim_{n \rightarrow \infty} \|f_n\|_{H_k} = \lim_{n \rightarrow \infty} \|f_n \circ A\|_{H_k} = \|f \circ A\|_{H_k}$$

We proceed to part 2. It is not hard to check that  $\rho$  is group homomorphism, so it only remains to validate that for every  $f \in H$  the mapping  $A \mapsto \rho(A)f$  is continuous. Let  $\epsilon > 0$  and let  $A \in \mathbb{O}(d)$ . We must show that there exists a neighbourhood  $U$  of  $A$  such that  $\forall B \in U$ ,  $\|f \circ A^{-1} - f \circ B^{-1}\|_{H_k} < \epsilon$ . Choose  $g(\cdot) = \sum_{i=1}^n \alpha_i k(\cdot, y_i)$  such that  $\|g - f\|_{H_k} < \frac{\epsilon}{3}$ . By part 1, it holds that

$$\begin{aligned} \|f \circ A^{-1} - f \circ B^{-1}\|_{H_k} &\leq \|f \circ A^{-1} - g \circ A^{-1}\|_{H_k} + \|g \circ A^{-1} - g \circ B^{-1}\|_{H_k} + \|g \circ B^{-1} - f \circ B^{-1}\|_{H_k} \\ &= \|f - g\|_{H_k} + \|g \circ A^{-1} - g \circ B^{-1}\|_{H_k} + \|g - f\|_{H_k} \\ &< \frac{\epsilon}{3} + \|g \circ A^{-1} - g \circ B^{-1}\|_{H_k} + \frac{\epsilon}{3} \end{aligned}$$

Thus, it is enough to find a neighbourhood  $U$  of  $A$  such that  $\forall B \in U, \|g \circ A^{-1} - g \circ B^{-1}\|_{H_k} < \frac{\epsilon}{3}$ . However,

$$\begin{aligned} \|g \circ A^{-1} - g \circ B^{-1}\|_{H_k}^2 &= \|g \circ A^{-1}\|_{H_k}^2 + \|g \circ B^{-1}\|_{H_k}^2 - 2 \operatorname{Re} \left[ \left\langle \sum_{i=1}^n \alpha_i k(\cdot, y_i) \circ A^{-1}, \sum_{i=1}^n \alpha_i k(\cdot, y_i) \circ B^{-1} \right\rangle \right] \\ &= 2\|g \circ A^{-1}\|_{H_k}^2 - 2 \operatorname{Re} \left[ \left\langle \sum_{i=1}^n \alpha_i k(\cdot, Ay_i), \sum_{i=1}^n \alpha_i k(\cdot, By_i) \right\rangle \right] \\ &= 2\|g \circ A^{-1}\|_{H_k}^2 - \operatorname{Re} \left[ \sum_{i,j=1}^n \alpha_i \bar{\alpha}_j k(By_j, Ay_i) \right] \end{aligned}$$

Since  $k$  is continuous, the last expression tends to  $2\|g \circ A^{-1}\|_{H_k}^2 - \operatorname{Re} \left[ \sum_{i,j=1}^n \alpha_i \bar{\alpha}_j k(Ay_j, Ay_i) \right] = \|g \circ A - g \circ A\|_{H_k}^2 = 0$  as  $B \rightarrow A$ . Thus, there exists a neighbourhood  $U$  such that  $\forall B \in U, \|g \circ A^{-1} - g \circ B^{-1}\|_{H_k} < \frac{\epsilon}{3}$  as required.

To see part 3, note that every function in  $H_k$  is a limit in  $H_k$  of functions of the form (9). Since  $k$  is continuous, every function in  $H_k$  is a limit in  $H_k$  of continuous functions. However, by Theorem 5.1, every function is in fact a uniform limit of continuous function, thus – continuous itself.

We proceed to part 4. By Theorem 5.4  $H_k = \oplus_{i \in I} V_i$  where each  $V_i$  is a finite dimensional space that is invariant to  $\rho$ . By Theorem 5.6 each  $V_i$  must be  $Y_n$  for some  $n$ , thus,  $H = \oplus_{n \in I} \mathbb{Y}_n^d$ . By the uniqueness part in Lemma 5.5 and Theorem 5.6, the restriction of  $\langle \cdot, \cdot \rangle_{H_k}$  to each  $\mathbb{Y}_n^d$ ,  $n \in I$  equals to  $\langle \cdot, \cdot \rangle_{L^2(S^{d-1})}$  up to scalar multiple, proving the formula for  $\langle \cdot, \cdot \rangle_{H_k}$ .

Finally, to see equation part 5, note that if for every  $n \in I$ ,  $\{Y_{n,j}^d\}_{j \in [N_{d,n}]}$  is an orthonormal basis of  $\mathbb{Y}_n^d$  w.r.t.  $\langle \cdot, \cdot \rangle_{L^2(S^{d-1})}$  then  $\{\frac{1}{a_n} Y_{n,j}^d\}_{n \in I, j \in [N_{d,n}]}$  is an orthogonal basis of  $H$ . By Theorem 5.3 and Lemma 5.7, it follows that, for every  $x \in S^{d-1}$ ,

$$1 = k(x, x) = \sum_{n \in I} a_n^{-2} \sum_{j=1}^{N_{d,n}} (Y_{n,j}^d(x))^2 = \sum_{n \in I} \frac{N_{d,n}}{|S^{d-1}|} a_n^{-2}$$

□

## Symmetrization

Let  $k : S^{d-1} \times S^{d-1} \rightarrow \mathbb{R}$  be a normalized continuous kernel. We define its *symmetrization* by

$$\forall x, y \in S^{d-1}, k_s(x, y) = \int_{\mathbb{O}(d)} k(Ax, Ay) dA$$

**Theorem 5.12** 1.  $k_s$  is symmetric continuous kernel with  $\sup_{x \in S^{d-1}} k_s(x, x) \leq 1$ .

2. For every  $\Phi \in L^2(\mathbb{O}(d), H_k)$  define  $\bar{\Phi} : S^{d-1} \rightarrow \mathbb{C}$  by  $\bar{\Phi}(x) = \int_{\mathbb{O}(d)} \Phi(A)(Ax) dA$ . Then

$$H_{k_s} = \{\bar{\Phi} : \Phi \in L^2(\mathbb{O}(d), H_k)\}$$

Moreover, for every  $\Phi \in L^2(\mathbb{O}(d), H_k)$ ,  $\|\bar{\Phi}\|_{H_{k_s}} \leq \|\Phi\|_{L^2(\mathbb{O}(d), H_k)}$ .

**Proof** Part 1. follows readily from the definition. We proceed to part 2. Define  $\phi : S^{d-1} \rightarrow L^2(\mathbb{O}(d), H_k)$  by

$$\phi(x)(A)(\cdot) = k(Ax, \cdot)$$

Note that

$$\begin{aligned} \langle \phi(x), \phi(y) \rangle_{L^2(\mathbb{O}(d), H_k)} &= \int_{\mathbb{O}(d)} \langle \phi(x)(A), \phi(y)(A) \rangle \\ &= \int_{\mathbb{O}(d)} \langle \phi(x)(A), \phi(y)(A) \rangle \\ &= k_s(x, y) \end{aligned}$$

Thus, the Theorem follows from Theorem 5.1.1

□

### 5.3 Lemma 5.16 and its proof

**Lemma 5.13** *For every  $n > 0, d \geq 5$  and  $t \in [-1, 1]$  it holds that*

$$|P_{d,n}(t)| \leq \min \left\{ \frac{\Gamma\left(\frac{d-1}{2}\right)}{\sqrt{\pi}} \left[ \frac{4}{n(1-t^2)} \right]^{\frac{d-2}{2}}, \left( \frac{n}{n+d-2} + 2|t| \right)^{\frac{n}{2}} \right\}$$

Moreover, if  $\frac{n}{n+d-2} + 2|t| \leq 1$  we also have

$$|P_{d,n}(t)| \leq \sqrt{\prod_{i=1}^n \left( \frac{i}{i+d-2} + 2|t| \right)}$$

Finally, there exist constants  $E > 0$  and  $0 < r, s < 1$  such that for every  $K > 0, d \geq 5$  and  $t \in [-\frac{1}{8}, \frac{1}{8}]$  we have

$$\sum_{n=K}^{\infty} |P_{d,n}(t)| \leq Er^K + Es^d$$

**Proof** In (Atkinson and Han, 2012) it is shown that  $|P_{d,n}(t)| \leq \frac{\Gamma\left(\frac{d-1}{2}\right)}{\sqrt{\pi}} \left[ \frac{4}{n(1-t^2)} \right]^{\frac{d-2}{2}}$ . We shall prove, by induction on  $k$  that

$$|P_{d,n}(t)| \leq \sqrt{\prod_{i=1}^n \left( \frac{i}{i+d-2} + 2|t| \right)}$$

Whenever  $\frac{n}{n+d-2} + 2|t| \leq 1$ . For  $n = 0, 1$  it follows from the fact that  $P_{d,0} \equiv 1$  and  $P_{d,1}(t) = t$ . Let  $n > 1$ . By the induction hypothesis and the recursion formula for the

Legendre polynomials we have

$$\begin{aligned}
|P_{d,n}(t)| &\leq \frac{2n+d-4}{n+d-3}|t||P_{d,n-1}(t)| + \frac{n-1}{n+d-3}|P_{d,n-2}(t)| \\
&\leq 2|t||P_{d,n-1}(t)| + \frac{n-1}{n+d-3}|P_{d,n-2}(t)| \\
&\leq 2|t|\sqrt{\prod_{i=1}^{n-1}\left(\frac{i}{i+d-2}+2|t|\right)} + \frac{n-1}{n+d-3}\sqrt{\prod_{i=1}^{n-2}\left(\frac{i}{i+d-2}+2|t|\right)} \\
&\leq 2|t|\sqrt{\prod_{i=1}^{n-2}\left(\frac{i}{i+d-2}+2|t|\right)} + \frac{n-1}{n+d-3}\sqrt{\prod_{i=1}^{n-2}\left(\frac{i}{i+d-2}+2|t|\right)} \\
&\leq \sqrt{\left(2|t|+\frac{n-1}{n+d-3}\right)\left(2|t|+\frac{n}{n+d-2}\right)}\sqrt{\prod_{i=1}^{n-2}\left(\frac{i}{i+d-2}+2|t|\right)} \\
&= \sqrt{\prod_{i=1}^n\left(\frac{i}{i+d-2}+2|t|\right)}
\end{aligned}$$

Now, every  $K, \bar{K} \geq 0$  such that  $\left(\frac{\bar{K}}{\bar{K}+d-2}+2|t|\right)^{\frac{1}{2}} < 1$ , we have

$$\begin{aligned}
\sum_{n=K}^{\infty} |P_{d,n}(t)| &\leq \sum_{n=K}^{\bar{K}} \left(\frac{n}{n+d-2}+2|t|\right)^{\frac{n}{2}} + \sum_{n=\bar{K}+1}^{\infty} \frac{\Gamma\left(\frac{d-1}{2}\right)}{\sqrt{\pi}} \left[\frac{4}{n(1-t^2)}\right]^{\frac{d-2}{2}} \\
&\leq \sum_{n=K}^{\bar{K}} \left(\frac{\bar{K}}{\bar{K}+d-2}+2|t|\right)^{\frac{n}{2}} + \sum_{n=\bar{K}+1}^{\infty} \frac{\Gamma\left(\frac{d-1}{2}\right)}{\sqrt{\pi}} \left[\frac{4}{n(1-t^2)}\right]^{\frac{d-2}{2}} \\
&\leq \sum_{n=K}^{\infty} \left(\frac{\bar{K}}{\bar{K}+d-2}+2|t|\right)^{\frac{n}{2}} + \frac{\Gamma\left(\frac{d-1}{2}\right)}{\sqrt{\pi}} \left[\frac{4}{(1-t^2)}\right]^{\frac{d-2}{2}} \sum_{n=\bar{K}+1}^{\infty} n^{-\frac{d-2}{2}} \\
&\leq \frac{\left(\frac{\bar{K}}{\bar{K}+d-2}+2|t|\right)^{\frac{K}{2}}}{1-\left(\frac{\bar{K}}{\bar{K}+d-2}+2|t|\right)^{\frac{1}{2}}} + \frac{\Gamma\left(\frac{d-1}{2}\right)}{\sqrt{\pi}} \left[\frac{4}{(1-t^2)}\right]^{\frac{d-2}{2}} \sum_{n=\bar{K}+1}^{\infty} n^{-\frac{d-2}{2}} \\
&\leq \frac{\left(\frac{\bar{K}}{\bar{K}+d-2}+2|t|\right)^{\frac{K}{2}}}{1-\left(\frac{\bar{K}}{\bar{K}+d-2}+2|t|\right)^{\frac{1}{2}}} + \frac{\Gamma\left(\frac{d-1}{2}\right)}{\sqrt{\pi}} \left[\frac{4}{(1-t^2)}\right]^{\frac{d-2}{2}} \int_{\bar{K}}^{\infty} x^{-\frac{d-2}{2}} dx \\
&= \frac{\left(\frac{\bar{K}}{\bar{K}+d-2}+2|t|\right)^{\frac{K}{2}}}{1-\left(\frac{\bar{K}}{\bar{K}+d-2}+2|t|\right)^{\frac{1}{2}}} + \frac{\Gamma\left(\frac{d-1}{2}\right)}{\sqrt{\pi}} \left[\frac{4}{(1-t^2)}\right]^{\frac{d-2}{2}} \frac{\bar{K}^{-\frac{d-4}{2}}}{\frac{d-4}{2}}
\end{aligned}$$

(We limit ourselves to  $d \geq 5$  to guarantee the convergence of  $\sum n^{-\frac{d-2}{2}}$ .) In particular, if  $|t| \leq \frac{1}{8}$  and  $\bar{K} = d - 2$ , we have,

$$\begin{aligned}
\sum_{n=K}^{\infty} |P_{d,n}(t)| &\leq \left( \frac{1}{1 - \left(\frac{3}{4}\right)^{\frac{1}{2}}} \right) \left( \frac{3}{4} \right)^{\frac{K}{2}} + \frac{\Gamma\left(\frac{d-1}{2}\right)}{\sqrt{\pi}} \left[ \frac{4.07}{(d-2)} \right]^{\frac{d-2}{2}} \frac{d-2}{\frac{d-4}{2}} \\
&\leq \left( \frac{1}{1 - \left(\frac{3}{4}\right)^{\frac{1}{2}}} \right) \left( \frac{3}{4} \right)^{\frac{K}{2}} + \frac{6\Gamma\left(\frac{d-1}{2}\right)}{\sqrt{\pi}} \left[ \frac{4.07}{(d-2)} \right]^{\frac{d-2}{2}} \\
&\sim \left( \frac{1}{1 - \left(\frac{3}{4}\right)^{\frac{1}{2}}} \right) \left( \frac{3}{4} \right)^{\frac{K}{2}} + \frac{6}{\sqrt{\pi}} \left[ \frac{4.07}{(d-2)} \right]^{\frac{d-2}{2}} \sqrt{\frac{2\pi}{2}} \left( \frac{d-2}{2e} \right)^{\frac{d-2}{2}} \\
&= \left( \frac{1}{1 - \left(\frac{3}{4}\right)^{\frac{1}{2}}} \right) \left( \frac{3}{4} \right)^{\frac{K}{2}} + 12 \left[ \frac{4.07}{2e} \right]^{\frac{d-2}{2}}
\end{aligned}$$

□

**Lemma 5.14** *Let  $\mu$  be a probability measure on  $[-1, 1]$  and let  $p_0, p_1, \dots$  be the corresponding orthogonal polynomials. Then, for every  $f \in \text{span}\{p_0, \dots, p_{K-1}\}$  we have*

$$\|f\|_2 \leq \sqrt{K} \|f\|_1 \cdot \max_{0 \leq i \leq K-1} \|p_i\|_{\infty}$$

Here, all  $L^p$  norms are w.r.t.  $\mu$ .

**Proof** Write  $f = \sum_{i=0}^{K-1} \alpha_i p_i$  and denote  $M = \max_{0 \leq i \leq K-1} \|p_i\|_{\infty}$ . We have

$$\begin{aligned}
\|f\|_2^2 &\leq \|f\|_1 \cdot \|f\|_{\infty} \\
&\leq \|f\|_1 \cdot M \sum_{n=0}^{K-1} |\alpha_k| \\
&\leq \|f\|_1 \cdot M \sqrt{\sum_{n=0}^{K-1} \alpha_k^2} \cdot \sqrt{K} \\
&= \|f\|_1 \cdot M \cdot \|f\|_2 \sqrt{K}
\end{aligned}$$

□

**Lemma 5.15** *Let  $d \geq 5$  and let  $f : [-1, 1] \rightarrow \mathbb{R}$  be a continuous function whose expansion in the basis of  $d$ -dimensional Legendre polynomials is*

$$f = \sum_{n=0}^{\infty} \alpha_n P_{d,n}$$

Denote  $C = \sup_n |\alpha_n|$ . Let  $\mu$  be the probability measure on  $[-1, 1]$  whose density function is

$$w(x) = \begin{cases} 0 & |x| > \frac{1}{8} \\ \frac{8}{\pi \sqrt{1-(8x)^2}} & |x| \leq \frac{1}{8} \end{cases}$$



Then, for every  $K \in \mathbb{N}$ ,  $\frac{1}{8} > \gamma > 0$ ,

$$|f(\gamma) - f(-\gamma)| \leq 32\gamma K^{3.5} \cdot \|f\|_{1,\mu} + (32\gamma K^{3.5} + 2) \cdot C \cdot E \cdot (r^K + s^d)$$

Here,  $E, r$  and  $s$  are the constants from Lemma 5.13.

**Proof** Let  $\bar{f} = \sum_{n=0}^{K-1} \alpha_n P_{d,n}$ . We have  $\|\bar{f}\|_{1,\mu} \leq \|f\|_{1,\mu} + \|\bar{f} - f\|_{\infty,\mu}$ . Define  $g : [-1, 1] \rightarrow \mathbb{R}$  by  $g(x) = \bar{f}(\frac{x}{8})$  and denote by  $d\lambda = \frac{dx}{\pi\sqrt{1-x^2}}$ . Write,

$$g = \sum_{n=0}^{K-1} \beta_n T_n$$

Where  $T_n$  are the Chebyshev polynomials. By Lemma 5.14 it holds that, for every  $0 \leq n \leq K-1$ ,

$$|\beta_n| \leq \sqrt{2} \|g\|_{2,\lambda} \leq 2\sqrt{K} \|g\|_{1,\lambda} = 2\sqrt{K} \|\bar{f}\|_{1,\mu}$$

Now,

$$g' = \sum_{n=1}^{K-1} \beta_n n U_{n-1}$$

Where  $U_n$  are the Chebyshev polynomials of the second kind. Thus,

$$\|g'\|_{\infty,\lambda} \leq \sum_{n=1}^{K-1} |\beta_n| \cdot n \cdot \|U_{n-1}\|_{\infty,\lambda} = \sum_{n=1}^{K-1} |\beta_n| \cdot n^2 \leq 2\sqrt{K} \|\bar{f}\|_{1,\mu} \cdot K^3$$

Finally, by Lemma 5.13,

$$\begin{aligned} |f(\gamma) - f(-\gamma)| &\leq |g(8\gamma) - g(-8\gamma)| + 2\|f - \bar{f}\|_{\infty,\mu} \\ &\leq 32\gamma K^{3.5} \cdot \|\bar{f}\|_{1,\mu} + 2\|f - \bar{f}\|_{\infty,\mu} \\ &\leq 32\gamma K^{3.5} \cdot (\|f\|_{1,\mu} + \|f - \bar{f}\|_{\infty,\mu}) + 2\|f - \bar{f}\|_{\infty,\mu} \\ &\leq 32\gamma K^{3.5} \cdot \|f\|_{1,\mu} + (32\gamma K^{3.5} + 2) \cdot \|f - \bar{f}\|_{\infty,\mu} \\ &\leq 32\gamma K^{3.5} \cdot \|f\|_{1,\mu} + (32\gamma K^{3.5} + 2) \cdot E \cdot C \cdot (r^K + s^d) \end{aligned}$$

□

For  $e \in S^{d-1}$  we define the group  $\mathbb{O}(e) := \{A \in \mathbb{O}(d) : Ae = e\}$ . If  $H_k$  be a symmetric RKHS and  $e \in S^{d-1}$  we define *Symmetrization around  $e$* . This is the operator  $\mathcal{P}_e : H_k \rightarrow H_k$  which is the projection on the subspace  $\{f \in H_k : \forall A \in \mathbb{O}(e), f \circ A = f\}$ . It is not hard to see that  $(\mathcal{P}_e f)(x) = \int_{\{x' : \langle x', e \rangle = \langle x, e \rangle\}} f(x') dx' = \int_{\mathbb{O}(e)} f \circ A(x) dA$ . Since  $\mathcal{P}_e f$  is a convex combination of the functions  $\{f \circ A\}_{A \in \mathbb{O}(e)}$ , it follows that if  $\mathcal{R} : H_k \rightarrow \mathbb{R}$  is a convex functional then  $\mathcal{R}(\mathcal{P}_e f) \leq \int_{\mathbb{O}(e)} \mathcal{R}(f \circ A) dA$ .

**Lemma 5.16 (main)** *There exists a probability measure  $\mu$  on  $[-1, 1]$  with the following properties. For every continuous and normalized kernel  $k : S^{d-1} \times S^{d-1} \rightarrow \mathbb{R}$  and  $C > 0$ , there exists  $e \in S^{d-1}$  such that, for every  $f \in H_k$  with  $\|f\|_{H_k} \leq C$ ,  $K \in \mathbb{N}$  and  $0 < \gamma < \frac{1}{8}$ ,*

$$\begin{aligned} \left| \int_{\{x : \langle x, e \rangle = \gamma\}} f - \int_{\{x : \langle x, e \rangle = -\gamma\}} f \right| &\leq 32\gamma K^{3.5} \cdot \|f\|_{1,\mu_e} + (32\gamma K^{3.5} + 2) \cdot E \cdot C \cdot (r^K + s^d) \\ &\leq 32\gamma K^{3.5} \cdot \|f\|_{1,\mu_e} + 10 \cdot E \cdot K^{3.5} \cdot C \cdot (r^K + s^d) \end{aligned}$$

The integrals are w.r.t. the uniform probability over  $\{x : \langle x, e \rangle = \gamma\}$  and  $\{x : \langle x, e \rangle = -\gamma\}$  and  $E, r, s$  are the constants from Lemma 5.13.

**Proof** Suppose first that  $k$  is symmetric. Let  $\mu$  be the distribution over  $[-1, 1]$  whose density function is

$$w(x) = \begin{cases} 0 & |x| > \frac{1}{8} \\ \frac{8}{\pi\sqrt{1-(8x)^2}} & |x| \leq \frac{1}{8} \end{cases}$$

We can assume that  $f$  is  $\mathbb{O}(e)$ -invariant. Otherwise, we can replace  $f$  with  $\mathcal{P}_e f$ , which does not change the l.h.s. and does not increase the r.h.s. This assumption yields (see (Atkinson and Han, 2012), pages 17-18)

$$f(x) = \sum_{n=0}^{\infty} \alpha_n P_{d,n}(\langle e, x \rangle).$$

The  $L^2(S^{d-1})$ -norm of the map  $x \mapsto P_{d,n}(\langle x, e \rangle)$  is  $\frac{|S^{d-1}|}{N_{d,n}}$  (e.g. (Atkinson and Han, 2012), page 71). Therefore,

$$\|f\|_k^2 = \sum_{n \in I} \frac{|S^{d-1}|}{N_{d,n}} a_n^2 \alpha_n^2$$

where  $\{a_n\}_{n \in I}$  are the numbers corresponding to  $H_k$  from Theorem 5.11. In particular (since also for  $n \notin I$ ,  $\alpha_n = 0$ ),

$$|\alpha_n|^2 \leq \frac{N_{d,n}}{|S^{d-1}|} a_n^{-2} \|f\|_k^2 \leq \|f\|_k^2$$

Write

$$g(t) = f(te), \quad t \in [-1, 1]$$

By Lemma 5.15,

$$|g(\gamma) - g(-\gamma)| \leq 32\gamma K^{3.5} \cdot \|f\|_{1,\mu} + (32\gamma K^{3.5} + 2) \cdot E \cdot C \cdot (r^K + s^d)$$

Finally,  $\int_{\{x: \langle x, e \rangle = \gamma\}} f = g(\gamma)$ ,  $\int_{\{x: \langle x, e \rangle = -\gamma\}} f = g(-\gamma)$  since  $f$  is  $\mathbb{O}(e)$ -invariant. The Lemma follows.

We proceed to the general case where  $k$  is not necessarily symmetric. Assume by way of contradiction that for every  $e \in S^{d-1}$ , there exists a function  $f_e$  such that

$$\int_{\{x: \langle x, e \rangle = \gamma\}} f_e - \int_{\{x: \langle x, e \rangle = -\gamma\}} f_e > 32\gamma K^{3.5} \cdot \|f_e\|_{1,\mu_e} + (32\gamma K^{3.5} + 2) \cdot \|f_e\|_{H_k} \cdot C \cdot (r^K + s^d) \quad (10)$$

For convenience we normalize, so l.h.s. equals 1. Fix a vector  $e_0 \in S^{d-1}$ . Define  $\Phi \in L^2(\mathbb{O}(d), H_k)$  by

$$\Phi(A) = f_{Ae_0}$$

and let  $f \in H_{k_s}$  be the function

$$f(x) = \int_{\mathbb{O}(d)} \Phi(A)(Ax) dA = \int_{\mathbb{O}(d)} f_{Ae_0}(Ax) dA$$

Now, it holds that

$$\begin{aligned}
\int_{\{x:\langle x, e_0 \rangle = \gamma\}} f - \int_{\{x:\langle x, e_0 \rangle = -\gamma\}} f &= \int_{\{x:\langle x, e_0 \rangle = \gamma\}} \int_{\mathbb{O}(d)} f_{Ae_0}(Ax) dA dx - \int_{\{x:\langle x, e_0 \rangle = -\gamma\}} \int_{\mathbb{O}(d)} f_{Ae_0}(Ax) dA dx \\
&= \int_{\mathbb{O}(d)} \int_{\{x:\langle x, e_0 \rangle = \gamma\}} f_{Ae_0}(Ax) dx - \int_{\{x:\langle x, e_0 \rangle = -\gamma\}} f_{Ae_0}(Ax) dx dA \\
&= \int_{\mathbb{O}(d)} \int_{\{x:\langle x, Ae_0 \rangle = \gamma\}} f_{Ae_0}(x) dx - \int_{\{x:\langle x, Ae_0 \rangle = -\gamma\}} f_{Ae_0}(x) dx dA \\
&= 1
\end{aligned}$$

On the other hand

$$\begin{aligned}
\|f\|_{1, \mu_e} &= \int_{S^{d-1}} \left| \int_{\mathbb{O}(d)} f_{Ae_0}(Ax) dA \right| d\mu_{e_0}(x) \\
&\leq \int_{\mathbb{O}(d)} \int_{S^{d-1}} |f_{Ae_0}(Ax)| d\mu_{e_0}(x) dA \\
&\leq \int_{\mathbb{O}(d)} \int_{S^{d-1}} |f_{Ae_0}(x)| d\mu_{Ae_0}(x) dA \\
&= \int_{\mathbb{O}(d)} \|f_{Ae_0}\|_{1, \mu_{Ae_0}} dA
\end{aligned}$$

Moreover, by Theorem 5.12,

$$\|f\|_{H_{k_s}}^2 \leq \|\Phi\|_{L^2(\mathbb{O}(d), H_k)}^2 = \int_{\mathbb{O}(d)} \|f_{Ae_0}\|_{H_k}^2 dA \leq C^2$$

Since the Lemma is already proved for symmetric kernels, it follows that

$$\begin{aligned}
1 &\leq 32\gamma K^{3.5} \cdot \|f\|_{1, \mu_{e_0}} + (32\gamma K^{3.5} + 2) \cdot E \cdot C \cdot (r^K + s^d) \\
&\leq 32\gamma K^{3.5} \cdot \int_{\mathbb{O}(d)} \|f_{Ae_0}\|_{1, \mu_{Ae_0}} dA + (32\gamma K^{3.5} + 2) \cdot E \cdot C \cdot (r^K + s^d) \\
&= \int_{\mathbb{O}(d)} 32\gamma K^{3.5} \cdot \|f_{Ae_0}\|_{1, \mu_{Ae_0}} + (32\gamma K^{3.5} + 2) \cdot E \cdot C \cdot (r^K + s^d) dA
\end{aligned}$$

Thus, for some  $A \in \mathbb{O}(d)$

$$1 \leq 32\gamma K^{3.5} \cdot \|f_{Ae_0}\|_{1, \mu_{Ae_0}} + (32\gamma K^{3.5} + 2) \cdot E \cdot C \cdot (r^K + s^d)$$

Contradicting Equation (10).

□

## 5.4 Proofs of the main Theorems

We are now ready to prove Theorems 2.5 and 3.2. We only consider distributions that supported on the unit sphere, and we can therefore assume that the problem is formulated

it terms of the unit sphere and not the unit ball. Also, we reformulate program (5) as follows: Given  $l : \mathbb{R} \rightarrow \mathbb{R}$  a convex surrogate, a constant  $C > 0$  and a continuous kernel  $k : S^\infty \times S^\infty \rightarrow \mathbb{R}$  with  $\sup_{x \in S^\infty} k(x, x) \leq 1$ , we want to solve

$$\begin{aligned} \min \quad & \text{Err}_{\mathcal{D}, l}(f + b) \\ \text{s.t.} \quad & f \in H_k, b \in \mathbb{R} \\ & \|f\|_{H_k} \leq C \end{aligned} \tag{11}$$

We can assume that  $\partial_+ l(0) < 0$ , for otherwise the approximation ratio is  $\infty$ . To see that, let the distribution  $\mathcal{D}$  be concentrated on a single point on the sphere and always return the label 1. Of course,  $\text{Err}_\gamma(\mathcal{D}) = 0$ . However, if  $\partial_+ l(0) \geq 0$ , it is not hard to see that if  $f, b$  is the solution of program (11), then  $f(x) + b \leq 0$ , so that  $\text{Err}_{0-1}(f + b) = 1$ .

**Lemma 5.17** *Let  $l$  be a surrogate loss,  $\mu$  a probability measure on  $S^{d-1}$  and  $f \in C(S^{d-1})$ . Let  $\bar{\mu}$  be the probability measure on  $S^{d-1} \times \{\pm 1\}$  which is the product measure of  $\mu$  and the uniform distribution on  $\{\pm 1\}$ . Then*

$$\|f\|_{1, \mu} \leq \frac{2}{|\partial_+ l(0)|} \text{Err}_{\bar{\mu}, l}(f)$$

**Proof** By Jansen's inequality, it holds that

$$\begin{aligned} \text{Err}_{\bar{\mu}, l}(f) &= \mathbb{E}_{(x, y) \sim \bar{\mu}} l(y \cdot f(x)) \\ &= \frac{1}{2} \mathbb{E}_{(x, y) \sim \bar{\mu}} l(f(x)) + l(-f(x)) \\ &\geq \frac{1}{2} \mathbb{E}_{(x, y) \sim \bar{\mu}} l(-|f(x)|) \\ &\geq \frac{1}{2} l(-\mathbb{E}_{(x, y) \sim \bar{\mu}} |f(x)|) \end{aligned}$$

It follows that  $l(-\|f\|_{1, \mu}) \leq 2 \text{Err}_{\bar{\mu}, l}(f)$ . By the convexity of  $l$ , it follows that for every  $x \in \mathbb{R}$ ,  $l(x) \geq l(0) + x \cdot \partial_+ l(0) = l(0) - x \cdot |\partial_+ l(0)| \geq -x \cdot |\partial_+ l(0)|$ . Thus,

$$\|f\|_{1, \mu} \leq \frac{2}{|\partial_+ l(0)|} \text{Err}_{\bar{\mu}, l}(f)$$

□

#### 5.4.1 Theorems 2.5 and 3.2

We will need Levy's measure concentration Lemma (e.g., (Milman and Schechtman, 2002)). Let  $f : X \rightarrow Y$  be an absolutely continuous map between metric spaces. We define its *modulus of continuity* as

$$\forall \epsilon > 0, \omega_f(\epsilon) = \sup\{d(f(x), f(y)) : x, y \in X, d(x, y) \leq \epsilon\}$$

**Theorem 5.18 (Levy's Lemma)** *There exists a constant  $\eta > 0$  such that for every continuous function  $f : S^{d-1} \rightarrow \mathbb{R}$ ,*

$$\Pr(|f - \mathbb{E}f| > \omega_f(\epsilon)) \leq \exp(-\eta d \epsilon^2)$$

*Here, both probability and expectation are w.r.t. the uniform distribution.*

We note that  $\omega_{f \circ g} \leq \omega_f \cdot \omega_g$  and that  $\omega_{\Lambda_v}(\epsilon) = \|v\| \cdot \epsilon$ . Thus, if  $\psi : S^\infty \rightarrow H_1$  is an absolutely continuous embedding such that  $k(x, y) = \langle \psi(x), \psi(y) \rangle_{H_1}$ , then for every  $v \in H_1$ , it holds that  $\omega_{\Lambda_{v,0} \circ \psi} \leq \|v\|_{H_1} \cdot \omega_\psi$ . Suppose now that  $f \in H_k$  with  $\|f\|_{H_k} \leq C$ . Let  $v \in H_1$  such that  $f = \Lambda_{v,0} \circ \psi$  and  $\|v\|_{H_1} = \|f\|_{H_k} \leq C$ . It follows from Levi's Lemma that

$$\Pr(|f - \mathbb{E}f| > C \cdot \omega_\psi(\epsilon)) \leq \Pr(|f - \mathbb{E}f| > \omega_f(\epsilon)) \leq \exp(-\eta d \epsilon^2) \quad (12)$$

Again, when both probability and expectation are w.r.t. the uniform distribution over  $S^{d-1}$ . **Proof** (of Theorems 2.5 and 3.2) Let  $\beta > \alpha > 0$  such that  $l(\alpha) > l(\beta)$ . Choose  $0 < \theta < 1$  large enough so that  $(1 - \theta)l(-\beta) + \theta l(\beta) < \theta l(\alpha)$ . Define probability measures  $\mu^1, \mu^2, \mu$  over  $[-1, 1] \times \{\pm 1\}$  as follows.

$$\mu^1((-\gamma, -1)) = 1 - \theta, \quad \mu^1((\gamma, 1)) = \theta$$

The measure  $\mu^2$  is the product of *uniform* $\{\pm 1\}$  and the measure on  $[-1, 1]$  whose density function is

$$w(x) = \begin{cases} 0 & |x| > \frac{1}{8} \\ \frac{8}{\pi \sqrt{1 - (8x)^2}} & |x| \leq \frac{1}{8} \end{cases}$$

Finally,  $\mu = (1 - \lambda)\mu^1 + \lambda\mu^2$  for  $\lambda > 0$ , which will be chosen later.

Let  $e \in S^{d-1}$  be the vector from Lemma 5.16. The distribution  $\mathcal{D}$  is the pullback of  $\mu$  w.r.t.  $e$ . By considering the affine functional  $\Lambda_{e,0}$ , it holds that  $\text{Err}_\gamma(\mathcal{D}) \leq \lambda$ .

Let  $g$  be the solution returned by the algorithm. With probability  $\geq 1 - \exp(-1/\gamma)$ ,  $g = f + b$ , where  $f, b$  is a solution to program (11) with  $C = C_A(\gamma)$  and with an additive error  $\leq \sqrt{\gamma}$ . Since the value of the zero solution for program (11) is  $l(0)$ , it follows that

$$l(0) + \sqrt{\gamma} \geq \text{Err}_{\mu,l}(g) = (1 - \lambda) \text{Err}_{\mu^1,l}(g) + \lambda \text{Err}_{\mu^2,l}(g)$$

Thus,  $\text{Err}_{\mu^2,l}(g) \leq \frac{l(0) + \sqrt{\gamma}}{\lambda} \leq \frac{2l(0)}{\lambda}$ . Combining Lemma 5.17 and Lemma 5.16 it follows that

$$\left| \int_{\{x: \langle x, e \rangle = \gamma\}} g - \int_{\{x: \langle x, e \rangle = -\gamma\}} g \right| \leq \frac{128l(0)\gamma K^{3.5}}{|\partial_+ l(0)|\lambda} + 10 \cdot K^{3.5} \cdot E \cdot C \cdot (r^K + s^d)$$

By choosing  $K = \Theta(\log(C))$ ,  $\lambda = \Theta(\gamma K^{3.5}) = \Theta(\gamma \log^{3.5}(C))$  and  $d = \Theta(\log(C))$ , we can make the last bound  $\leq \frac{\alpha}{2}$ . We claim that  $\int_{\{x: \langle x, e \rangle = -\gamma\}} g > \frac{\alpha}{2}$ . To see that, note that otherwise  $\int_{\{x: \langle x, e \rangle = \gamma\}} g \leq \alpha$  thus,

$$\begin{aligned} \mathbb{E}_{(x,y) \sim \mathcal{D}} l((f(x) + b)y) &= \mathbb{E}_{(x,y) \sim \mathcal{D}} l(g(x)y) \\ &\geq \theta(1 - \lambda) \cdot \int_{\{x: \langle x, e \rangle = \gamma\}} l(g(x)) dx \\ &\geq \theta(1 - \lambda) \cdot l \left( \int_{\{x: \langle x, e \rangle = \gamma\}} g(x) dx \right) \\ &\geq \theta \cdot l(\alpha) \cdot (1 - \lambda) = \theta \cdot l(\alpha) + o(1) \end{aligned}$$

This contradicts the optimality of  $f, b$ , as for  $f' = 0, b' = \beta$  it holds that

$$\begin{aligned} \mathbb{E}_{(x,y) \sim \mathcal{D}} l((f'(x) + b')y) &\leq \lambda l(-\beta) + (1 - \lambda) \cdot (1 - \theta)l(-\beta) + \theta \cdot l(\beta) \\ &= (1 - \theta)l(-\beta) + \theta \cdot l(\beta) + o(1) \end{aligned}$$

We can conclude now the proof of Theorem 2.5. By choosing  $d$  large enough and using Equation (12), we can guarantee that  $g|_{\{x:\langle x,e \rangle = -\gamma\}}$  is very concentrated around its expectation. In particular, if  $(x, y)$  are sampled according to  $\mathcal{D}$ , then w.p.  $> 0.5 \cdot (1 - \theta) \cdot (1 - \lambda) = \Omega(1)$ , it holds that  $yg(x) < 0$ . Thus,  $\text{Err}_{\mathcal{D},0-1}(g) = \Omega(1)$ , while  $\text{Err}_\gamma(\mathcal{D}) \leq \lambda = O(\gamma \text{poly}(\log(C)))$

To conclude the proof of Theorem 3.2, we note that we can assume that  $g$  is  $\mathbb{O}(e)$ -invariant. Otherwise, we can replace it with  $\mathcal{P}_e f + b$ . This does not increase  $\|f\|_{H_k}$  nor  $\text{Err}_{\mathcal{D},l}(f + b)$ , thus, the solution  $\mathcal{P}_e f + b$  is optimal as well. Now, it follows that  $g|_{\{x:\langle x,e \rangle = -\gamma\}}$  is constant and we finish as before.  $\square$

#### 5.4.2 Theorem 3.1

Let  $L$  be the Lipschitz constant of  $l$ . Let  $\beta > \alpha > 0$  such that  $l(\alpha) > l(\beta)$ . Choose  $0 < \theta < 1$  large enough so that  $(1 - \theta)l(-\beta) + \theta l(\beta) < \theta l(\alpha)$ . First, define probability measures  $\mu^1, \mu^2, \mu^3$  and  $\mu$  over  $[-1, 1] \times \{\pm 1\}$  as follows.

$$\begin{aligned}\mu^1(\gamma, 1) &= \theta, \quad \mu^1(-\gamma, -1) = 1 - \theta \\ \mu^2(-\gamma, 1) &= 1\end{aligned}$$

The measure  $\mu^3$  is the product of  $\text{uniform}\{\pm 1\}$  and the measure over  $[-1, 1]$  whose density function is

$$w(x) = \begin{cases} 0 & |x| > \frac{1}{8} \\ \frac{8}{\pi \sqrt{1 - (8x)^2}} & |x| \leq \frac{1}{8} \end{cases}$$

Finally,  $\mu = (1 - \lambda_1 - \lambda_2)\mu^1 + \lambda_2\mu^2 + \lambda_3\mu^3$  with  $\lambda_2, \lambda_3 > 0$  to be chosen later.

Now, let  $e \in S^{d-1}$  be the vector from Lemma 5.16. The distribution  $\mathcal{D}$  is the pullback of  $\mu$  w.r.t.  $e$ . By considering the affine functional  $\Lambda_{e,0}$ , it holds that  $\text{Err}_\gamma(\mathcal{D}) \leq \lambda_3 + \lambda_2$ .

Let  $g$  be the solution returned by the algorithm. With probability  $\geq 1 - \exp(-1/\gamma)$ ,  $g = f + b$ , where  $f, b$  is a solution to program (11) with  $C = C_A(\gamma)$  and with an additive error  $\leq \sqrt{\gamma}$ . As in the proof of Theorem 2.5, it holds that

$$\left| \int_{\{x:\langle x,e \rangle = \gamma\}} g - \int_{\{x:\langle x,e \rangle = -\gamma\}} g \right| \leq \frac{128l(0)\gamma K^{3.5}}{|\partial_+ l(0)|\lambda_3} + 10 \cdot K^{3.5} \cdot E \cdot C \cdot (r^K + s^d) \quad (13)$$

Denote the last bound by  $\epsilon$ . It holds that

$$\text{Err}_{\mathcal{D},l}(g) = (1 - \lambda_2 - \lambda_3)\mathbb{E}_{\mu_e^1}l(yg(x)) + \lambda_2\mathbb{E}_{\mu_e^2}l(yg(x)) + \lambda_3\mathbb{E}_{\mu_e^3}l(yg(x)) \quad (14)$$

Now, denote  $\delta = \int_{\{x:\langle x,e \rangle = -\gamma\}} g$ . It holds that

$$\begin{aligned}\mathbb{E}_{\mu_e^1}l(yg(x)) &= \theta \int_{\{x:\langle x,e \rangle = \gamma\}} l(g(x)) + (1 - \theta) \int_{\{x:\langle x,e \rangle = -\gamma\}} l(-g(x)) \\ &\geq \theta \cdot l\left(\int_{\{x:\langle x,e \rangle = \gamma\}} g\right) + (1 - \theta) \cdot l\left(-\int_{\{x:\langle x,e \rangle = -\gamma\}} g\right) \\ &\geq \theta \cdot l(\delta) + (1 - \theta) \cdot l(-\delta) - L\epsilon\end{aligned} \quad (15)$$

Thus,

$$\text{Err}_{\mathcal{D},l}(g) \geq (1 - \lambda_2 - \lambda_3)(\theta \cdot l(\delta) + (1 - \theta) \cdot l(-\delta)) - L\epsilon + \lambda_2 \mathbb{E}_{\mu_e^2} l(yg(x))$$

However, by considering the constant solution  $\delta$ , it follows that

$$\begin{aligned} \text{Err}_{\mathcal{D},l}(g) &\leq (1 - \lambda_2 - \lambda_3)(\theta l(\delta) + (1 - \theta) \cdot l(-\delta)) + \lambda_2 \cdot l(\delta) + \lambda_3 \frac{1}{2} (l(\delta) + l(-\delta)) + \sqrt{\gamma} \\ &\leq (1 - \lambda_2 - \lambda_3)(\theta \cdot l(\delta) + (1 - \theta) \cdot l(-\delta)) + \lambda_2 \cdot l(\delta) + \lambda_3 \cdot l(-|\delta|) + \sqrt{\gamma} \end{aligned}$$

Thus,

$$\begin{aligned} \text{Err}_{\mu_e^2,l}(g) &\leq \frac{L\epsilon}{\lambda_2} + l(\delta) + \frac{\lambda_3}{\lambda_2} l(-|\delta|) + \frac{\sqrt{\gamma}}{\lambda_2} \\ &= \frac{L \cdot l(0) 128 \gamma K^{3.5}}{|\partial_+ l(0)| \lambda_2 \lambda_3} + \frac{10 \cdot L \cdot K^{3.5}}{\lambda_2} \cdot E \cdot C \cdot (r^K + s^d) + l(\delta) + \frac{\lambda_3}{\lambda_2} l(-|\delta|) + \frac{\sqrt{\gamma}}{\lambda_2} \end{aligned} \quad (16)$$

Now, relying on the assumption that  $\gamma \cdot \log^8(C) = o(1)$ , it is possible to choose  $\lambda_2 = \Theta(\sqrt{\gamma} K^4) = \Theta(\sqrt{\gamma} \log^4(C))$ ,  $\lambda_3 = \sqrt{\gamma}$ ,  $K = \Theta(\log(C/\gamma))$ , and  $d = \Theta(\log(C/\gamma))$  such that the bound in Equation (13),  $\frac{L \cdot l(0) 128 \gamma K^{3.5}}{|\partial_+ l(0)| \lambda_2 \lambda_3} + \frac{10 K^{3.5}}{\lambda_2} \cdot E \cdot C \cdot (r^K + s^d)$ ,  $\lambda_2$ ,  $\lambda_3$  and  $\frac{\lambda_3}{\lambda_2}$  are all  $o(1)$ .

Since the bound in Equation (13) is  $o(1)$ , it follows, as in the proof of Theorem 2.5, that  $l(\delta) \leq l(\frac{\alpha}{2})$  and consequently,  $0 < \frac{\alpha}{2} \leq \delta$ . From equations (14) and (15), it follows that

$$l(-|\delta|) = l(-\delta) \leq \frac{L\epsilon + \frac{\text{Err}_{\mathcal{D},l}(g)}{1 - \lambda_2 - \lambda_3}}{1 - \theta} \leq \frac{L\epsilon + \frac{2l(0)}{1 - \lambda_2 - \lambda_3}}{1 - \theta} = O(1)$$

It now follows from Equation (16) that

$$\mathbb{E}_{(x,y) \sim \mu_2} l(g(x)y) = \text{Err}_{\mu_e^2,l}(g) \leq l\left(\frac{\alpha}{2}\right) + o(1)$$

By Markov's inequality,

$$\Pr_{(x,y) \sim \mu_2} (l(g(x)y) \geq l(0)) \leq \frac{l\left(\frac{\alpha}{2}\right) + o(1)}{l(0)}$$

Thus, if  $(x, y)$  are chosen according to  $\mu_e^2$ , then w.p.  $> \frac{l(0) - l(\frac{\alpha}{2})}{l(0)} - o(1)$ ,  $l(g(x)) < l(0) \Rightarrow g(x) > 0$ . Since the marginal distributions of  $\mu_e^1$  and  $\mu_e^2$  are the same, it follows that, if  $(x, y)$  are chosen according to  $\mathcal{D}$ , then w.p.  $> \left( \frac{l(0) - l(\frac{\alpha}{2})}{l(0)} - o(1) \right) \cdot (1 - \lambda_2 - \lambda_3) \cdot (1 - \theta) = \Omega(1)$ ,  $yg(x) < 0$ . Thus,  $\text{Err}_{\mathcal{D},0-1}(g) = \Omega(1)$  while  $\text{Err}_{\gamma}(\mathcal{D}) \leq \lambda_2 + \lambda_3 = O(\sqrt{\gamma} \text{poly}(\log(C)))$ .

□



### 5.4.3 The integrality gap – Theorem 3.3

Our first step is a reduction to the hinge loss. Let  $a = \partial_+ l(0)$ . Define

$$l^*(x) = \begin{cases} ax + 1 & x \leq \frac{1}{a} \\ 0 & \text{o/w} \end{cases}$$

it is not hard to see that  $l^*$  is a convex surrogate satisfying  $\forall x, l^*(x) \leq l(x)$  and  $\partial_+ l^*(0) = \partial_+ l(0)$ . Thus, if we substitute  $l$  with  $l^*$ , we just decrease the integrality gap, hence can assume that  $l = l^*$ . Now, we note that if we consider program (11) with  $l = l^*$  the integrality gap coincides with what we get by replacing  $C$  with  $|a| \cdot C$  and  $l^*$  with the hinge loss. To see that, note that for every  $f \in H_k, b \in \mathbb{R}$ ,  $\text{Err}_{\mathcal{D}, l^*}(f + b) = \text{Err}_{\mathcal{D}, \text{hinge}}(|a| \cdot f + |a| \cdot b)$ , thus, minimizing  $\text{Err}_{\mathcal{D}, l^*}$  over all functions  $f \in H_k$  that satisfy  $\|f\|_{H_k} \leq C$  is equivalent to minimizing  $\text{Err}_{\mathcal{D}, \text{hinge}}$  over all functions  $f \in H_k$  that satisfy  $\|f\|_{H_k} \leq |a| \cdot C$ . Thus, it is enough to prove the Theorem for  $l = l_{\text{hinge}}$ .

Next, we show that we can assume that the embedding is symmetric (i.e., correspond to a symmetric kernel). As the integrality gap is at least as large as the approximation ratio, using Theorem 3.2 this will complete our argument. (The reduction to the hinge loss yields bounds with universal constants in the asymptotic terms).

Let  $\gamma > 0$  and let  $\mathcal{D}$  be a distribution on  $S^{d-1} \times \{\pm 1\}$ . It is enough to find (a possibly different) distribution  $\mathcal{D}_1$  with the same  $\gamma$ -margin error as  $\mathcal{D}$ , for which the optimum of program (11) (with  $l = l_{\text{hinge}}$ ) is not smaller than the optimum of the program

$$\begin{aligned} \min \quad & \text{Err}_{\mathcal{D}, \text{hinge}}(f + b) \\ \text{s.t.} \quad & f \in H_{k_s}, b \in \mathbb{R} \\ & \|f\|_{H_{k_s}} \leq C \end{aligned} \tag{17}$$

Denote the optimal value of program (17) by  $\alpha$  and assume, towards contradiction, that whenever  $\text{Err}_\gamma(\mathcal{D}_1) = \text{Err}_\gamma(\mathcal{D})$ , the optimum of program (11) is strictly less than  $\alpha$ .

For every  $A \in \mathbb{O}(d)$ , let  $\mathcal{D}_A$  be the distribution of the r.v.  $(Ax, y) \in S^{d-1} \times \{\pm 1\}$ , where  $(x, y) \sim \mathcal{D}$ . Since clearly  $\text{Err}_\gamma(\mathcal{D}_A) = \text{Err}_\gamma(\mathcal{D})$ , there exist  $f_A \in H_k$  and  $b_A \in \mathbb{R}$  such that  $\|f_A\|_{H_k} \leq C$  and  $\text{Err}_{\mathcal{D}_A, \text{hinge}}(g_A) < \alpha$ , where  $g_A := f_A + b_A$ . Define  $f \in H_{k_s}$  by  $f(x) = \int_{\mathbb{O}(d)} f_A(Ax) dA$  and let  $b = \int_{\mathbb{O}(d)} b_A dA$  and  $g = f + b$ . By Theorem 5.12,  $\|f\|_{H_{k_s}} \leq C$ . Finally, for  $l = l_{\text{hinge}}$ ,

$$\begin{aligned} \text{Err}_{\mathcal{D}, \text{hinge}}(g) &= \mathbb{E}_{(x, y) \sim \mathcal{D}} l(yg(x)) \\ &= \mathbb{E}_{(x, y) \sim \mathcal{D}} l(y \mathbb{E}_{A \sim \mathbb{O}(d)} g_A(Ax)) \\ &\leq \mathbb{E}_{(x, y) \sim \mathcal{D}} \mathbb{E}_{A \sim \mathbb{O}(d)} l(yg_A(Ax)) \\ &= \mathbb{E}_{A \sim \mathbb{O}(d)} \mathbb{E}_{(x, y) \sim \mathcal{D}} l(yg_A(Ax)) \\ &= \mathbb{E}_{A \sim \mathbb{O}(d)} \mathbb{E}_{(x, y) \sim \mathcal{D}_A} l(yg_A(x)) < \alpha \end{aligned}$$

Contrary to the assumption that  $\alpha$  is the optimum of program (17).

### 5.4.4 Finite dimension - Theorems 2.6 and 3.4

Let  $V \subseteq C(S^{d-1})$  be the linear space  $\{\Lambda_{v,b} \circ \psi : v \in \mathbb{R}^m, b \in \mathbb{R}\}$  and denote  $\bar{W} = \{\Lambda_{v,b} \circ \psi : v \in W, b \in \mathbb{R}\}$ . We note that  $\dim(V) \leq m + 1$ . Instead of program (4) we consider the

equivalent formulation

$$\begin{aligned} \min \quad & \text{Err}_{\mathcal{D},l}(f) \\ \text{s.t.} \quad & f \in \bar{W} \end{aligned} \tag{18}$$

**Lemma 5.19 (John's Lemma)** (*Matousek, 2002*) *Let  $V$  be an  $m$ -dimensional real vector space and let  $K$  be a full-dimensional 0-symmetric compact convex set. Then there exists an inner product on  $V$  so that  $K$  is contained in the unit ball, and contains the ball around 0 of radius  $\frac{1}{\sqrt{m}}$ .*

**Lemma 5.20** *Let  $l$  be a convex surrogate and let  $V \subset C(S^{d-1})$  an  $m$ -dimensional vector space. There exists a continuous kernel  $k : S^{d-1} \times S^{d-1} \rightarrow \mathbb{R}$  with  $\sup_{x \in S^{d-1}} k(x, x) \leq 1$  such that  $H_k = V$  as a vector space and there exists a probability measure  $\mu_N$  such that*

$$\forall f \in V, \|f\|_{H_k} \leq \frac{2m^{1.5}}{|\partial_+ l(0)|} \text{Err}_{\mu_N, l}(f)$$

**Proof** Let  $\psi : S^{d-1} \rightarrow V^*$  be the evaluation operator. It maps each  $x \in S^{d-1}$  to the linear functional  $f \in V \mapsto f(x)$ . We claim that

1.  $\psi$  is continuous,
2.  $\text{aff}(\psi(S^{d-1}) \cup -\psi(S^{d-1})) = V^*$ ,
3.  $V = \{v^{**} \circ \psi : v^{**} \in V^{**}\}$ .

Proof of **1**: We need to show that  $\psi(x_n) \rightarrow \psi(x)$  if  $x_n \rightarrow x$ . Since  $V^*$  is finite dimensional, it suffices to show that  $\psi(x_n)(f) \rightarrow \psi(x)(f)$  for every  $f \in V$ , which follows from the continuity of  $f$ .

Proof of **2**: Note that  $0 \in U = \text{aff}(\psi(S^{d-1}) \cup -\psi(S^{d-1}))$ , so  $U$  is a linear space. Now, define  $T : U^* \rightarrow V$  via  $T(u^*) = u^* \circ \psi$ . We claim that  $T$  is onto, whence  $\dim(U) = \dim(U^*) = \dim(V) = \dim(V^*)$ , so that  $U = V^*$ . Indeed, for  $f \in V$ , let  $u_f^* \in U^*$  be the functional  $u_f^*(u) = u(f)$ . Now,  $T(u_f^*)(x) = u_f^*(\psi(x)) = \psi(x)(f) = f(x)$ , thus  $T(u_f^*) = f$ .

Proof of **3**: From  $U = V^*$  it follows that  $U^* = V^{**}$ , so that the mapping  $T : V^{**} \rightarrow V$  is onto, showing that  $V = \{v^{**} \circ \psi : v^{**} \in V^{**}\}$ .

Let us apply John's Lemma to  $K = \text{conv}(\psi(S^{d-1}) \cup -\psi(S^{d-1}))$ . It yields an inner product on  $V^*$  with  $K$  contained in the unit ball and containing the ball around 0 with radius  $\frac{1}{\sqrt{m}}$ . Let  $k$  be the kernel  $k(x, y) = \langle \psi(x), \psi(y) \rangle$ . Since  $\psi$  is continuous,  $k$  is continuous as well. By Theorem 5.1.1 and since  $T$  is onto, it follows that, as a vector space,  $V = H_k$ . Since  $K$  is contained in the unit ball, it follows that  $\sup_{x \in S^{d-1}} k(x, x) \leq 1$ . It remains to prove the existence of the measure  $\mu_N$ .

Let  $e_1, \dots, e_m \in V^*$  be an orthonormal basis. For every  $i \in [m]$ , choose  $(x_i^1, y_i), \dots, (x_i^{m+1}, y_i) \in S^{d-1} \times \{\pm 1\}$  and  $\lambda_i^1, \dots, \lambda_i^{m+1} \geq 0$  such that  $\sum_{j=1}^{m+1} \lambda_i^j = 1$  and  $\frac{1}{\sqrt{m}} e_i = \sum_{j=1}^{m+1} \lambda_i^j y_i \psi(x_i^j)$ . Define  $\mu_N(x_i^j, 1) = \mu_N(x_i^j, -1) = \frac{\lambda_i^j}{2m}$ .

Let  $f \in V$ . By Theorem 5.1.1 there exists  $v \in V^*$  such that  $f = \Lambda_{v,0} \circ \psi$  and  $\|f\|_{H_k} = \|v\|_{V^*}$ . It follows that, for  $a = \partial_+ l(0)$ ,

$$\begin{aligned}
\text{Err}_{\mu_N, l}(f) &= \sum_{i=1}^m \sum_{j=1}^{m+1} \frac{\lambda_i^j}{2m} [l(y_i f(x_i^j)) + l(-y_i f(x_i^j))] \\
&\geq \frac{1}{2m} \sum_{i=1}^m \left[ l \left( \sum_{j=1}^{m+1} \lambda_i^j y_i f(x_i^j) \right) + l \left( - \sum_{j=1}^{m+1} \lambda_i^j y_i f(x_i^j) \right) \right] \\
&= \frac{1}{2m} \sum_{i=1}^m \left[ l \left( \sum_{j=1}^{m+1} \lambda_i^j y_i \langle v, \psi(x_i^j) \rangle \right) + l \left( - \sum_{j=1}^{m+1} \lambda_i^j y_i \langle v, \psi(x_i^j) \rangle \right) \right] \\
&= \frac{1}{2m} \sum_{i=1}^m l \left( \langle v, \sum_{j=1}^{m+1} \lambda_i^j y_i \psi(x_i^j) \rangle \right) + l \left( - \langle v, \sum_{j=1}^{m+1} \lambda_i^j y_i \psi(x_i^j) \rangle \right) \\
&= \frac{1}{2m} \sum_{i=1}^m l \left( \langle v, \frac{e_i}{\sqrt{m}} \rangle \right) + l \left( - \langle v, \frac{e_i}{\sqrt{m}} \rangle \right) \\
&\geq \frac{1}{2m} \sum_{i=1}^m l \left( - \frac{|\langle v, e_i \rangle|}{\sqrt{m}} \right) \\
&\geq \frac{|a|}{2m^{1.5}} \sum_{i=1}^m |\langle v, e_i \rangle| \\
&\geq \frac{|a|}{2m^{1.5}} \|v\|_{V^*} = \frac{|a|}{2m^{1.5}} \|f\|_{H_k}
\end{aligned}$$

□

**Proof** (of Theorem 2.6) Let  $L$  be the Lipschitz constant of  $l$ . Let  $\beta > \alpha > 0$  such that  $l(\alpha) > l(\beta)$ . Choose  $0 < \theta < 1$  large enough so that  $(1 - \theta)l(-\beta) + \theta l(\beta) < \theta l(\alpha)$ . First, define probability measures  $\mu^1, \mu^2, \mu^3$  and  $\mu$  over  $[-1, 1] \times \{\pm 1\}$  as follows.

$$\mu^1(\gamma, 1) = \theta, \quad \mu^1(-\gamma, -1) = 1 - \theta$$

$$\mu^2(-\gamma, 1) = 1$$

The measure  $\mu^3$  is the product of  $\text{uniform}\{\pm 1\}$  and the measure over  $[-1, 1]$  whose density function is

$$w(x) = \begin{cases} 0 & |x| > \frac{1}{8} \\ \frac{8}{\pi \sqrt{1 - (8x)^2}} & |x| \leq \frac{1}{8} \end{cases}$$

Let  $k, \mu_N$  be the distribution and kernel from Lemma 5.20. Now, let  $e \in S^{d-1}$  be the vector from Lemma 5.16. We define the distribution  $\mathcal{D}$  corresponding to the measure

$$\mu = (1 - \lambda_2 - \lambda_3 - \lambda_N) \mu_e^1 + \lambda_2 \mu_e^2 + \lambda_3 \mu_e^3 + \lambda_N \mu_N$$

By considering the affine functional  $\Lambda_{e,0}$ , it holds that  $\text{Err}_\gamma(\mathcal{D}) \leq \lambda_3 + \lambda_2 + \lambda_N$ .

Let  $g$  be the solution returned by the algorithm. With probability  $\geq 1 - \exp(-1/\gamma)$ ,  $g = f + b$ , where  $f, b$  is a solution to program (18) with an additive error  $\leq \sqrt{\gamma}$ .

Denote  $\|g\|_{H_k} = C$ . By Lemma 5.20, it holds that

$$\begin{aligned} C &\leq \frac{2m^{1.5}}{|\partial_+ l(0)|} \text{Err}_{\mu_N, l}(g) \\ &\leq \frac{2m^{1.5}}{|\partial_+ l(0)|} \frac{\text{Err}_{\mu, l}(g)}{\lambda_N} \\ &\leq \frac{2m^{1.5}}{|\partial_+ l(0)|} \frac{l(0)}{\lambda_N} \end{aligned}$$

As in the proof of Theorem 2.5, it holds that

$$\left| \int_{\{x: \langle x, e \rangle = \gamma\}} g - \int_{\{x: \langle x, e \rangle = -\gamma\}} g \right| \leq \frac{128l(0)\gamma K^{3.5}}{|\partial_+ l(0)|\lambda_3} + 10 \cdot K^{3.5} \cdot E \cdot C \cdot (r^K + s^d) \quad (19)$$

Denote the last bound by  $\epsilon$ . It holds that

$$\text{Err}_{\mathcal{D}, l}(g) = (1 - \lambda_2 - \lambda_3 - \lambda_N) \mathbb{E}_{\mu_e^1} l(yg(x)) + \lambda_2 \mathbb{E}_{\mu_e^2} l(yg(x)) + \lambda_3 \mathbb{E}_{\mu_e^3} l(yg(x)) + \lambda_N \mathbb{E}_{\mu_N} l(yg(x)) \quad (20)$$

Now, denote  $\delta = \int_{\{x: \langle x, e \rangle = -\gamma\}} g$ . It holds that

$$\begin{aligned} \mathbb{E}_{\mu_e^1} l(yg(x)) &= \theta \int_{\{x: \langle x, e \rangle = \gamma\}} l(g(x)) + (1 - \theta) \int_{\{x: \langle x, e \rangle = -\gamma\}} l(-g(x)) \\ &\geq \theta \cdot l \left( \int_{\{x: \langle x, e \rangle = \gamma\}} g \right) + (1 - \theta) \cdot l \left( - \int_{\{x: \langle x, e \rangle = -\gamma\}} g \right) \\ &\geq \theta \cdot l(\delta) + (1 - \theta) \cdot l(-\delta) - L\epsilon \end{aligned} \quad (21)$$

Thus,

$$\text{Err}_{\mathcal{D}, l}(g) \geq (1 - \lambda_2 - \lambda_3 - \lambda_N)(\theta \cdot l(\delta) + (1 - \theta) \cdot l(-\delta)) - L\epsilon + \lambda_2 \mathbb{E}_{\mu_e^2} l(yg(x))$$

However, by considering the constant solution  $\delta$ , it follows that

$$\begin{aligned} \text{Err}_{\mathcal{D}, l}(g) &\leq (1 - \lambda_2 - \lambda_3 - \lambda_N)(a \cdot l(\delta) + (1 - \theta) \cdot l(-\delta)) + \lambda_2 \cdot l(\delta) + (\lambda_3 + \lambda_N) \frac{1}{2} (l(\delta) + l(-\delta)) + \sqrt{\gamma} \\ &\leq (1 - \lambda_2 - \lambda_3 - \lambda_N)(\theta \cdot l(\delta) + (1 - \theta) \cdot l(-\delta)) + \lambda_2 \cdot l(\delta) + (\lambda_3 + \lambda_N) \cdot l(-|\delta|) + \sqrt{\gamma} \end{aligned}$$

Thus,

$$\begin{aligned} \text{Err}_{\mu_e^2, l}(g) &\leq \frac{L\epsilon}{\lambda_2} + l(\delta) + \frac{\lambda_3 + \lambda_N}{\lambda_2} l(-|\delta|) + \frac{\sqrt{\gamma}}{\lambda_2} \\ &\leq \frac{L \cdot l(0) 128 \gamma K^{3.5}}{|\partial_+ l(0)| \lambda_2 \lambda_3} + \frac{10 \cdot L \cdot K^{3.5}}{\lambda_2} \cdot E \cdot C \cdot (r^K + s^d) + l(\delta) + \frac{\lambda_3 + \lambda_N + \sqrt{\gamma}}{\lambda_2} l(-|\delta|) \end{aligned} \quad (22)$$

Now, relying on the assumption that  $\gamma \cdot \log^8(C) = o(1)$ , it is possible to choose  $\lambda_2 = \Theta(\sqrt{\gamma} K^4) = \Theta(\sqrt{\gamma} \log^4(C))$ ,  $\lambda_3 = \sqrt{\gamma}$ ,  $K = \Theta(\log(C/\gamma))$ ,  $\lambda_N = \gamma$  and  $d = \Theta(\log(C/\gamma))$

such that the bound in Equation (19),  $\frac{L \cdot l(0) 128 \gamma K^{3.5}}{|\partial_+ l(0)| \lambda_2 \lambda_3} + \frac{10 K^{3.5}}{\lambda_2} \cdot E \cdot C \cdot (r^K + s^d)$ ,  $\lambda_2$ ,  $\lambda_3$ ,  $\lambda_N$  and  $\frac{\lambda_3 + \lambda_N + \sqrt{\gamma}}{\lambda_2}$  are all  $o(1)$ .

Since the bound in Equation (19) is  $o(1)$ , it follows, as in the proof of Theorem 2.5, that  $l(\delta) \leq l\left(\frac{\alpha}{2}\right)$  and consequently,  $0 < \frac{\alpha}{2} \leq \delta$ . From equations (20) and (21), it follows that

$$l(-|\delta|) = l(-\delta) \leq \frac{L\epsilon + \frac{\text{Err}_{\mathcal{D},l}(g)}{1-\lambda_2-\lambda_3-\lambda_N}}{1-\theta} \leq \frac{L\epsilon + \frac{2l(0)}{1-\lambda_2-\lambda_3-\lambda_N}}{1-\theta} = O(1)$$

It now follows from Equation (22) that

$$\mathbb{E}_{(x,y) \sim \mu_2} l(g(x)y) = \text{Err}_{\mu_e^2, l}(g) \leq l\left(\frac{\alpha}{2}\right) + o(1)$$

By Markov's inequality,

$$\Pr_{(x,y) \sim \mu_2} (l(g(x)y) \geq l(0)) \leq \frac{l\left(\frac{\alpha}{2}\right) + o(1)}{l(0)}$$

Thus, if  $(x, y)$  are chosen according to  $\mu_e^2$ , then w.p.  $> \frac{l(0) - l\left(\frac{\alpha}{2}\right)}{l(0)} - o(1)$ ,  $l(g(x)) < l(0) \Rightarrow g(x) > 0$ . Since the marginal distributions of  $\mu_e^1$  and  $\mu_e^2$  are the same, it follows that, if  $(x, y)$  are chosen according to  $\mathcal{D}$ , then w.p.  $> \left(\frac{l(0) - l\left(\frac{\alpha}{2}\right)}{l(0)} - o(1)\right) \cdot (1 - \lambda_2 - \lambda_3 - \lambda_N) \cdot (1 - \theta) = \Omega(1)$ ,  $yg(x) < 0$ . Thus,  $\text{Err}_{\mathcal{D}, 0-1}(g) = \Omega(1)$  while  $\text{Err}_{\gamma}(\mathcal{D}) \leq \lambda_2 + \lambda_3 + \lambda_N = O(\sqrt{\gamma} \text{poly}(\log(C))) = O(\sqrt{\gamma} \text{poly}(\log(m/\gamma)))$ .

□

**Proof** (of Theorem 3.4) As in the proof of Theorem 3.3, we can assume w.l.o.g. that  $l = l_{\text{hinge}}$ . Let  $k, \mu_N$  be the measure and the kernel from Lemma 5.20. Let  $C = 2m^{1.5}/\gamma$ . By (the proof of) Theorem 3.3, there exists a probability measure  $\bar{\mu}$  over  $S^{d-1} \times \{\pm 1\}$  such that for every  $f \in H_k$  with  $\|f\|_{H_k} \leq C$  it holds that  $\text{Err}_{\bar{\mu}, l}(f) = \Omega(1)$  but  $\text{Err}_{\gamma}(\bar{\mu}) = O(\gamma \cdot \text{poly}(\log(C)))$ . Consider the distribution  $\mu = (1 - \gamma)\bar{\mu} + \gamma\mu_N$ . It still holds that  $\text{Err}_{\gamma}(\bar{\mu}) = O(\gamma \cdot \text{poly}(\log(C))) = O(\gamma \cdot \text{poly}(\log(m/\gamma)))$ . Let  $f$  be an optimal for program (18). We have that  $1 \geq \text{Err}_{\mu, l}(f) \geq \gamma \cdot \text{Err}_{\mu_N, l}(f)$ . By Lemma 5.20,  $\|f\|_{H_k} \leq C$ . Thus,  $\text{Err}_{\mu, l}(f) \geq (1 - \gamma) \text{Err}_{\bar{\mu}, l}(f) = \Omega(1)$ .

□

## 6 Choosing a surrogate according to the margin.

The purpose of this section is to demonstrate the subtleties relating to the possibility of choosing a convex surrogate  $l$  according to the margin  $\gamma$ . Let  $k : B \times B \rightarrow \mathbb{R}$  be the kernel

$$k(x, y) = \frac{1}{1 - \frac{1}{2} \langle x, y \rangle_H}$$

and let  $\psi : B \rightarrow H_1$  be a corresponding embedding (i.e.,  $k(x, y) = \langle \psi(x), \psi(y) \rangle_{H_1}$ ). In (Shalev-Shwartz et al., 2011) it has been shown that the solution  $f, b$  to Program (2), with  $C = C(\gamma) = \text{poly}(\exp(1/\gamma \cdot \log(1/\gamma)))$  and the embedding  $\psi$ , satisfies

$$\text{Err}_{\text{hinge}}(f + b) \leq \text{Err}_{\gamma}(\mathcal{D}) + \gamma .$$

Consequently, every approximated solution to the Program with an additive error of at most  $\gamma$  will have a 0-1 loss bounded by  $\text{Err}_{\gamma}(\mathcal{D}) + 2\gamma$ .

For every  $\gamma$ , define a 1-Lipschitz convex surrogate by

$$l_{\gamma}(x) = \begin{cases} 1 - x & x \leq 1/C(\gamma) \\ 1 - 1/C(\gamma) & x \geq 1/C(\gamma) \end{cases}$$

**Claim 1** *A function  $g : B \rightarrow \mathbb{R}$  is a solutions to Program (5) with  $l = l_{\gamma}$ ,  $C = 1$  and the embedding  $\psi$ , if and only if  $C(\gamma) \cdot g$  is a solutions to Program (2) with  $C = C(\gamma)$  and the embedding  $\psi$ .*

We postpone the proof to the end of the section. We note that Program (5) with  $l = l_{\gamma}$ ,  $C = 1$  and the embedding  $\psi$ , have a complexity of 1, according to our conventions. Moreover, by Claim 1, the optimal solution to it has a 0-1 error of at most  $\text{Err}_{\gamma}(\mathcal{D}) + \gamma$ . Thus, if  $A$  is an algorithm that is only obligated to return an approximated solution to Program (5) with  $l = l_{\gamma}$ ,  $C = 1$  and the embedding  $\psi$ , we cannot lower bound its approximation ratio. In particular, our Theorems regarding the approximation ratio are no longer true, as currently stated, if the algorithms are allowed to choose the surrogate according to  $\gamma$ . One might be tempted to think that by the above construction (i.e. taking  $\psi$  as our embedding, choosing  $C = 1$  and  $l = l_{\gamma}$ , and approximate the program upon a sample of size  $\text{poly}(1/\gamma)$ ), we have actually gave 1-approximation algorithm. The crux of the matter is that algorithms that approximate the program according to a finite sample of size  $\text{poly}(1/\gamma)$  are only guaranteed to find a solution with an additive error of  $\text{poly}(\gamma)$ . For the loss  $l_{\gamma}$ , such an additive error is meaningless: Since for every function  $f$ ,  $\text{Err}_{\mathcal{D}, l_{\gamma}}(f) \geq 1 - 1/C(\gamma)$ , the 0 solution has an additive error of  $\text{poly}(\gamma)$ . Therefore, we cannot argue that the solution returned by the algorithm will have a small 0-1 error. Indeed we anticipate that the algorithm we have described will suffer from serious over-fitting.

To summarize, we note that the lower bounds we have proved, relies on the fact that the optimal solutions of the programs we considered are very bad. For the algorithm we sketched above, the optimal solution is very good. However, guaranties on approximated solutions obtained from a polynomial sample are meaningless. We conclude that lower bounds for such algorithms will have to involve over-fitting arguments, which are out of the scope of the paper.

**Proof** (of claim 1) Define

$$l_{\gamma}^*(x) = \begin{cases} 1 - C(\gamma)x & x \leq \frac{1}{C(\gamma)} \\ 0 & x \geq \frac{1}{C(\gamma)} \end{cases}$$

Since  $l_\gamma^*(x) = C(\gamma) \cdot (l_\gamma(x) - (1 - \frac{1}{C(\gamma)}))$ , it follows that the solutions to Program (5) with  $l = l_\gamma^*$ ,  $C = 1$  and  $\psi$  coincide with the solutions with  $l = l_\gamma$ ,  $C = 1$  and  $\psi$ . Now, we note that, for every function  $f : B \rightarrow \mathbb{R}$ ,

$$\text{Err}_{\mathcal{D}, l_\gamma^*}(f) = \text{Err}_{\mathcal{D}, \text{hinge}}(C(\gamma) \cdot f)$$

Thus,  $w, b$  minimizes  $\text{Err}_{\mathcal{D}, l_\gamma^*}(\Lambda_{w,b} \circ \psi)$  under the restriction that  $\|w\| \leq 1$  if and only if  $C(\gamma) \cdot w, C(\gamma) \cdot b$  minimizes  $\text{Err}_{\mathcal{D}, \text{hinge}}(\Lambda_{w,b} \circ \psi)$  under the restriction that  $\|w\| \leq C(\gamma)$ .

□

## References

- Martin Anthony and Peter Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.
- K. Atkinson and W. Han. *Spherical Harmonics and Approximations on the Unit Sphere: An Introduction*, volume 2044. Springer, 2012.
- P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006.
- S. Ben-David and H. Simon. Efficient learning of linear perceptrons. In *NIPS*, 2000.
- S. Ben-David, D. Loker, N. Srebro, and K. Sridharan. Minimizing the misclassification error rate using a surrogate convex loss. In *ICML*, 2012.
- A. Birnbaum and S. Shalev-Shwartz. Learning halfspaces with the zero-one loss: Time-accuracy tradeoffs. In *NIPS*, 2012.
- E. Blais, R. O'Donnell, and K. Wimmer. Polynomial regression under arbitrary product distributions. In *COLT*, 2008.
- S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.
- N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.
- V. Feldman, P. Gopalan, S. Khot, and A.K. Ponnuswami. New results for learning noisy parities and halfspaces. In *In Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, 2006.
- G.B. Folland. *A course in abstract harmonic analysis*. CRC, 1994.
- V. Guruswami and P. Raghavendra. Hardness of learning halfspaces with noise. In *Proceedings of the 47th Foundations of Computer Science (FOCS)*, 2006.

- A. Kalai, A.R. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. In *Proceedings of the 46th Foundations of Computer Science (FOCS)*, 2005.
- Kosaku Yosida. *Functional Analysis*. Springer-Verlag, Heidelberg, 1963.
- P.M. Long and R.A. Servedio. Learning large-margin halfspaces with more malicious noise. In *NIPS*, 2011.
- J. Matousek. *Lectures on discrete geometry*, volume 212. Springer, 2002.
- V.D. Milman and G. Schechtman. *Asymptotic Theory of Finite Dimensional Normed Spaces: Isoperimetric Inequalities in Riemannian Manifolds*, volume 1200. Springer, 2002.
- S. Saitoh. *Theory of reproducing kernels and its applications*. Longman Scientific & Technical England, 1988.
- IJ Schoenberg. Positive definite functions on spheres. *Duke. Math. J.*, 1942.
- B. Schölkopf, C. Burges, and A. Smola, editors. *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1998.
- S. Shalev-Shwartz, O. Shamir, and K. Sridharan. Learning kernel-based halfspaces with the 0-1 loss. *SIAM Journal on Computing*, 40:1623–1646, 2011.
- I. Steinwart and A. Christmann. *Support vector machines*. Springer, 2008.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58(1):267–288, 1996.
- V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.
- T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32:56–85, 2004.